Development of Rainfall Forecast Performance Monitoring Criteria

Phase 1: Development of Methodology and Algorithms

A E Jones, D A Jones and R J Moore

CEH Wallingford
Maclean Building
Crowmarsh Gifford
Wallingford
Oxfordshire
OX10 8BB
UK

Commissioning Organisation

Environment Agency
Rivers House
Waterside Drive
Aztec West
Bristol BS32 4UD
Tel: +44(0)1454 624400
Fax: +44(0)1454 624409

Met Office
London Road
Bracknell
Berkshire
RG12 2SY
Tel: +44(0)1344 855680
Fax: +44(0)1344 855681

Dissemination status
Internal: Released Internally
External: Released to Public Domain.

Statement of Use
This document describes the findings of a Project that aimed to develop rainfall forecast performance monitoring criteria for use by the Met Office and the Environment Agency.

Contractor
This document was produced under Met Office Contract Number PB/B3734 by:

CEH Wallingford
Maclean Building
Crowmarsh Gifford
Wallingford, Oxon
OX10 8BB
Tel: +44 (0) 1491 838800 Fax: +44 (0) 1491 692424

Project Leader
The Project Leader for Contract PB/B3734 was:
Bryony May – Joint Centre for Hydro-Meteorological Research, Wallingford
The Environment Agency Client was:
Alison Pickles – National Flood Warning Centre

Further copies of this report are available from:
CEH Wallingford, Maclean Building, Crowmarsh Gifford, Wallingford, Oxon, OX10 8BB, UK
Tel: 01491-838800 Fax: 01491-692424 e-mail: rm@ceh.ac.uk

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

ACKNOWLEDGEMENTS

Particular thanks are due to the following members of the Project Steering Committee:

Met Office:
Bryony May (Project Manager and Committee Chairman)
Dave Smith
Environment Agency:
Alison Pickles

CONTENTS

LIST OF TABLES AND FIGURES

EXECUTIVE SUMMARY

Criteria for monitoring the quality of rainfall forecasts employed in support of flood warning are required to assess their reliability in use and to provide feedback aimed at providing an improved service. The rainfall forecasts of main concern here are the quantitative component of the Daily Weather Forecast, the Evening Update and the Heavy Rainfall Warning. These three products are produced by the Met Office as a service to the Environment Agency in support of their flood warning responsibilities.

This report, commissioned jointly by the Environment Agency and the Met Office, first reviews current methodology and practice in monitoring the performance of rainfall forecasts. The content, format and delivery mechanisms of each of the three forecast products are also reviewed and recommendations for revision made. The report proceeds to develop a framework for assessment, addressing issues such as selection of performance measures, choice of "ground truth", and sources of comparative forecasts such as rainfall forecasts obtained directly from the Mesoscale Model and from the Nimrod radar-based product. New methods for assessing the accuracy of performance measures - as determined by a given rainfall forecast, ground-truth and comparative forecast dataset -are introduced.

Rainfall forecasts for case study storms are used to trial the assessment procedure employing a selection of performance measures. The case study storms were chosen by the Environment Agency to be of flooding interest to a number of its regions. Suitable ground truth available for assessment, including raingauge and Nimrod quality-controlled radar data, are identified and processed to a form suitable for application in the analysis. The analysis of the case study dataset is used to develop practical experience in the use of the assessment procedure leading to recommendations for operational implementation. These recommendations concern both the automated assessment of forecasts and the use of a PC tool with manual data-entry for assessing the Heavy Rainfall Warnings. The development of the PC tool features as an important operational output of the project.

The report concludes with a summary of the study, encompassing its main conclusions and recommendations. In particular, this points out the advantages of using a small and rather simple set of performance measures. The mean absolute error provides an easily understood and stable measure of the "typical size of error", in the same units as the rainfall forecast. For a categorical measure of rainfall threshold exceedence, the Critical Success Index and False Alarm Rate provide a useful pairing that are widely used and easily understood. For assessing probability forecasts, the Continuous Brier Score provides a simple measure analogous in form to the mean absolute error. Measures of forecast bias are also included in the selected set of performance measures considered important.

KEYWORDS

Rainfall, forecast, performance, assessment, verification

1. INTRODUCTION

1.1 Background Requirement

The broad aim of this study was to develop an objective means of assessing the performance of the Met Office rainfall forecasts used to support the issuing of Flood Watches and Flood Warnings by the Environment Agency. Within this broad remit a more specific aim was to establish performance criteria to be applied to the Daily Weather Forecasts, to the Evening Updates and to the Heavy Rainfall Warnings. It is these three Met Office forecast products that the Environment Agency currently rely on for information on future rainfall, complemented by radar-based forecasts out to 6 hours ahead.

The study also sought to review the content, format and delivery methods associated with these three forecast products, limited to the rainfall information they contain. This review aimed to fully appreciate the Agency user requirement for rainfall forecasts (automated flood forecasting, setting triggers, informal uses,…). It was also to consider the capability of the state-of-the-art of rainfall forecasting to provide better information via an efficient and timely delivery mechanism. It was to be expected that the textual information content of these products will continue to have value to the Agency at an informal level. However, the opportunity exists for improved levels of quantitative information about future rainfall (including its uncertainty) from Numerical Weather Prediction (NWP) models, where spatial resolution is becoming more refined, and for an enhanced automated delivery of their forecasts. This might argue for more radical changes to the forecast products, at least in the longer term. It was thus seen as an important part of this study to encompass such considerations when developing methodology and algorithms for performance assessment.

The study saw the comparison of the existing Daily Weather Forecasts with the NWP mesoscale model forecasts of rainfall as being of fundamental importance. If NWP forecasts outperform the Daily Weather Forecasts in the quantitative prediction of areal rainfall totals, then a more radical review of the Daily Weather Forecast product may need to be sought as it has no “added value” in this component. The study has therefore sought an assessment methodology that identifies such added value (positive or negative).

A key issue to be addressed was the formulation of an assessment framework which employs a “ground truth” that fairly judges the performance of rainfall forecasts presented as intensity ranges in an interval, as probabilities, and with respect to prescribed areas and local extremes. Raingauge (point) and radar (grid) information are available operationally to construct the “ground truth” estimates for the areas concerned.

An important operational output of the study was to be a simple PC-based facility to support application of the assessment methodology as far as it can be implemented via manual data entry, and assuming that the rainfall forecast products remain largely unchanged. The PC facility was to place emphasis on the Heavy Rainfall Warnings, as it was envisaged that the algorithms used to assess the Daily Weather Forecasts and Evening Updates will eventually run as part of a routine, automated process. The PC system development was to be simple, and recognise that the main focus of the study was the development of a rainfall forecast performance assessment methodology with a detailed consideration of different options.

1.2 Outline of the report

The report is made up of reviews of the methodologies for assessment and the existing rainfall forecast products, and the development of a methodology for forecast assessment and its application to rainfall forecasts and ground truths for case study storms. This leads to a summary and conclusion final section containing the main recommendations of the study.

The review of methodologies for assessing the performance of rainfall forecasts is presented in Section 2, focussing on the advantages and disadvantages of different Assessment Measures and the procedures in current use within the Environment Agency and the Met Office. The content, format and method of delivery of the present rainfall forecast products received by the Agency – the Daily Weather Forecast, Evening Updates and Heavy Rainfall Warnings – are reviewed in Section 3, and recommendations for improvement made.

A procedure for assessing the performance of the rainfall forecasts is developed in Section 4. This includes consideration of the choice of "ground truth" and how to assess the accuracy of performance measures, as determined by a given dataset of forecasts and ground-truths. Section 5 applies the assessment procedure to each of the three rainfall forecast products, for a selection of rainfall events identified by the Environment Agency. Forecasts of rainfall accumulations and rates are assessed, in single-value, category (single-value exceedence) and probability form. Section 5 is long and detailed and may be skipped over for a busy reader or on a first reading of the report. The results it contains provide important justifications for the conclusions that follow.

Against the experience gained from the case study assessment, recommendations are made in Section 6 on the form of assessment to use operationally, including choice of ground truths and performance measures. Section 6 also encompasses a summary and the main conclusions arising from the study.

2. REVIEW OF ASSESSMENT METHODOLOGIES

2.1 Introduction

The aim of this overall section is to review methodologies available for assessment of rainfall forecasts. Section 2.2 focuses on a review of performance measures available to assess the quality of forecasts. This is followed by a review of methodologies for assessing rainfall forecasts, paying special attention to those in use by the Environment Agency and the Met Office. A summary of recommendations arising from the review concludes the Section.

2.2 Review of Assessment Measures

2.2.1 Introduction

There are several good existing reviews of forecast assessment methods, and the assessment measures employed within them, for application in the hydro-meteorological sciences. For a concise review see Chapter 7 on *Forecast Verification* in the book by Wilks (1995) entitled *Statistical methods in the atmospheric sciences. An Introduction*. A classical and much referenced work commissioned by the World Meteorological Office is the *Survey of common verification methods in meteorology* written by Stanski, Wilson and Burrows (1995). Only just published in March 2003 is the book *Forecast Verification. A practioner's Guide in Atmospheric Science* edited by Jolliffe and Stephenson which provides an excellent and comprehensive state-of-the art account suited to both forecast practioners and researchers. The Reference section of this Report includes a bibliography containing a selection of the more important publications consulted as part of this review.

It would be wrong to produce a similar review here. Instead, an attempt has been made to summarise the wealth of information available on Assessment Measures into two tables. The first, presented as Table 2.2.1, gives definitions of the main Assessment Measures in precise form as mathematical formulae. This is complemented by Table 2.2.2 which provides a simple verbal description of each measure, the symbol (usually an acronym) used to represent it, the range of values it can take (and an indication of the best), and most importantly a summary of its advantages and disadvantages as an Assessment Measure. These tables are discussed further below.

2.2.2 Notation

Table 2.2.1 introduces a set of notation that is used consistently in this report with reference to the Assessment Measures. The symbol $\hat{y}_i$ is used to denote the i'th of a set of n rainfall forecasts whilst y_i denotes its "observed" value derived from some "ground-truth". Thus a simple scalar (additive) *error* is defined as $e_i = y_i - \hat{y}_i$, and a *log-error,* defined as

$$e_i^L = -\ln(\hat{y}_i / y_i) = \ln(y_i) - \ln(\hat{y}_i)$$

which deflates the error for larger rainfalls. An *error factor* may be defined as $f_i = \hat{y}_i / y_i$, a *proportional error* as $\varepsilon_i = (y_i - \hat{y}_i) / y_i$ and a *percentage error* as 100 times this.

Table 2.2.1 Assessment Measures: (a) Formulae for Continuous Measures

Assessment Measure	Formula
Basic quantities	y_i is the observed value of rainfall for sample i (i=1,2..., n). $\hat{y}_i$ is the forecast value of rainfall for sample i. $z_i = \log^* y_i$ is the observed value of "log-rainfall", where $\log^*$ is a revised version of the logarithm which gives a valid value for zero-rainfall. The definition of the revised logarithm used here is $\log^* x = \ln x$ for $x \geq \alpha$ or $\ln(\alpha/2)$ for $x < \alpha$; α is 0.2 when y_i is a rain amount in mm, and is 0.8 when y_i is a rain rate in mm h^{-1}. Alternative definitions are possible. $\hat{z}_i = \log^* \hat{y}_i$ is forecast value of log-rainfall
Bias (mean error)	$bias = n^{-1}\sum(y_i - \hat{y}_i)$
Mean error of log-rainfall	$mel = n^{-1}\sum(z_i - \hat{z}_i)$
Mean absolute error	$mae = n^{-1}\sum\lvert y_i - \hat{y}_i\rvert$
Mean absolute error of log-rainfall	$mael = n^{-1}\sum\lvert z_i - \hat{z}_i\rvert$
Mean square error	$mse = n^{-1}\sum(y_i - \hat{y}_i)^2$
Root mean square error	$rmse = \sqrt{n^{-1}\sum(y_i - \hat{y}_i)^2}$
Root mean square error of log-rainfall	$rmsel = \sqrt{n^{-1}\sum(z_i - \hat{z}_i)^2}$
Root Mean Square Factor	$rmsf = \exp\left\{\frac{1}{n}\sum\left[\ln\left(\frac{\hat{y}_i}{y_i}\right)\right]^2\right\}^{1/2}$ For practical applications this formula needs to be revised to avoid problems with the logarithm of 0. Two distinct possibilities exist which produce basically different results: (a) In the above formula, a revised-logarithm (defined under "basic quantities") might be used. (b) In the above formula, certain terms are omitted entirely from the summation. One version of this (Golding, 1998) is as follows. The summation only includes samples satisfying either $\alpha < \hat{y}_i < \beta$ or $\alpha < y_i < \beta$, to suppress the effect of trivial forecasts, and n is revised to be the number of terms in the summation actually used. Possible parameter values are α=0.2 mm and $\beta = \infty$. Further, the values used in the summation are replaced by trimmed versions: $\hat{y}_i^* = \max(\hat{y}, \alpha/2)$ $\quad \hat{y}_i^* = \min(\hat{y}_i, 2\beta)$ $y_i^* = \max(y_i, \alpha/2)$ $\quad y_i^* = \min(y_i, 2\beta)$ to avoid the logarithm of 0 and to place an upper limit on any one sample's contribution to the overall error.

Table 2.2.1 cont' Assessment Measures: (a) Formulae for Continuous Measures

R^2 (Efficiency)	$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$ $\bar{y} = n^{-1} \sum y_i$ is the sample mean of the observations.
R^2 for performance measures	(i) R^2 for *rmse*: R^2_{rmse} is the same as R^2 above. This can also be written as $R^2_{rmse} = 1 - \frac{mse}{mse_0}$ where mse_0 is the smallest mean square error obtained by any constant-valued forecast, in which case the forecast is equal to the sample mean. (ii) R^2 for *mae:* $R^2_{mae} = 1 - \frac{mae}{mae_0}$, with $mae_0 = n^{-1} \sum \mid y_i - \tilde{y} \mid$, the minimum value for a constant forecast of the mean absolute error with $\tilde{y}$ the median of y_i. (iii) R^2 for *rmsel*: R^2_{rmsel} as R^2 above but replacing y with z (iii) R^2 for *mael*: $R^2_{mael} = 1 - \frac{mael}{mael_0}$, with $mael_0 = n^{-1} \sum \mid z_i - \tilde{z} \mid$, the minimum value for a constant forecast of the mean absolute error of log-rainfall with $\tilde{z}$ the median of z_i.
Skill Score (generic)	$SS = \frac{P - P_{ref}}{P_{perf} - P_{ref}}$ P Performance of forecast according to chosen Performance Measure P_{ref} Performance of reference forecast P_{perf} Performance of perfect forecast
Correlation coefficient	$r = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}_i)}{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}}_i)^2}$ $\bar{\hat{y}} = n^{-1} \sum \hat{y}_i$ is sample mean of forecasts.

Table 2.2.1 cont' Assessment Measures: (b) Categorical Skill Scores

Categorical Skill Scores

Contingency table:

Event Forecast	Event Observed		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Ordinary Scores	
Hit Rate (Proportion Correct)	$H = \frac{a+d}{a+b+c+d}$
Critical Success Index (Threat Score)	$CSI = \frac{a}{a+b+c}$
False Alarm Rate	$FAR = \frac{b}{a+b}$
Probability of Detection (Hit Rate for observed 'yes')	$POD = \frac{a}{a+c}$
Probability of False Detection	$PFD = \frac{b}{b+d}$
CSI:POD:FAR Relation	$CSI = \frac{1}{\frac{1}{POD} + \frac{1}{1-FAR} - 1}$
Bias Ratio	$B = \frac{a+b}{a+c}$
Relative Scores	
Heidke Skill Score	$HSS = \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)}$
Kuipers Skill Score (Peirce's)	$KSS = \frac{(ad-bc)}{(a+c)(b+d)} = POD - PFD$
Equitable Threat Score (Gilbert Skill Score, GSS)	$ETS = \frac{(ad-bc)}{(a+b+c+d)(b+c)+(ad-bc)}$

Table 2.2.1 cont' Assessment Measures: (b) Categorical Skill Scores

Likelihood Ratio	LR_1 is the Likelihood Ratio for correct forecast of "below X" $LR_1 = \frac{d(a+c)}{c(b+d)}$ LR_2 is the Likelihood Ratio for correct forecast of "above X" $LR_2 = \frac{a(b+d)}{b(a+c)}$
Odds Ratio	$\theta = \frac{ad}{bc} = LR_1 LR_2$
Likelihood Ratio Benefit	LRB_1 is the Likelihood Ratio Benefit for correct forecast of "below X" $LRB_1 = \frac{LR_1}{LR_1^{ref}}$ LRB_2 is the Likelihood Ratio Benefit for correct forecast of "above X" $LRB_2 = \frac{LR_2}{LR_2^{ref}}$ LR_1^{ref}, LR_2^{ref} are the Likelihood Ratios for a reference forecast
Odds Ratio Benefit	$\theta_{ref}^{+} = \frac{\theta}{\theta_{ref}} = LRB_1 LRB_2$ θ_{ref} is the Odds Ratio for a reference forecast $\theta_{ref} = 1$ for "climatology" or "independence" when the Odds Ratio Benefit equates to the Odds Ratio.

Table 2.2.1 cont' Assessment Measures: (c) Skill Scores for Probability Forecasts

Categorical	
Brier Score	$BS = n^{-1}\sum (Y_i - \hat{Y}_i)^2$ Y_i indicator of event $y_i \le x$ in the observed sample, equal to 1 if event $y_i \le x$ does occur, 0 if not $\hat{Y}_i$ probability of event $y_i \le x$ occurring , as stated in the probability forecast, value in the range 0 to 1 Here y_i is the observed value of sample *i* (*i*=1,2..., *n*), and *x* is a threshold value defining the categories of event-occurrence or non-occurrence. A Multiple-Category Brier Score (MCBS) exists as an extension of this for probability forecasts for *k* event thresholds.
Brier Skill Score	$BSS = 1 - \frac{BS}{BS_{ref}}$ BS_{ref} is the Brier Score for a reference forecast (eg. climatological relative frequencies)
Continuous	
Continuous Brier Score	$BS = n^{-1}\sum \int (Y_i(x) - \hat{Y}_i(x))^2 dx$ Y_i (x) indicator of event $y_i \le x$ in the observed sample, equal to 1 if event $y_i \le x$ does occur, 0 if not $\hat{Y}_i$ (x) probability of event $y_i \le x$ occurring, as stated in the probability forecast, value in the range 0 to 1 Here y_i is the observed value of sample *i* (*i*=1,2..., *n*), and *x* is a variable threshold value covering all possible values of rainfall amount or rate. (A Continuous Brier Skill Score can be defined similarly to BSS above)

Table 2.2.2 Overview of Assessment Measures

Assessment Measure	**Symbol**	**Range of values** (Best value is indicated by *)	**Description**	**Advantages**	**Disadvantages**
Bias (mean error)	*bias*	-∞ to 0* to ∞	Mean of the errors.	Gives clear indication of forecast bias in units of forecast quantity.	Can suppress size of typical errors. Relates to an additive adjustment of the forecast, which is inappropriate to rainfall forecasts because of the special nature of "zero rainfall".
Mean error of log-rainfall	*mel*	-∞ to 0* to ∞	Mean of the errors of log-rainfall.	Gives clear indication of size of forecast error as a factor of observed value. Appropriate if errors are proportional to rainfall values being estimated.	Interpretation obscured by need to use revised logarithm.
Mean absolute error	*mae*	0* to ∞	Mean of the absolute values of the errors.	Gives typical size of error, independent of sign.	Masks effect of forecast bias and its sign. Less sensitive than *rmse* to large errors.
Mean absolute error of log-rainfall	*mael*	0* to ∞	Mean of the absolute values of the errors of log-rainfall.	Gives typical relative size of error as a factor of observations, independent of direction. Appropriate if errors are proportional to rainfall value being estimated.	Masks effect of forecast error bias and its sign. Less sensitive than *rmsel* to large errors. Notionally, implies zero error for zero rainfall, which is clearly untrue. Actual interpretation obscured by use of revised logarithm.
Mean square error	*mse*	0* to ∞	Mean of the squared errors.	Useful as a component of other summary statistics and as a quantity often used in practical/ theoretical statistics in more general comparisons of predictions and outcomes.	Not directly on a useful scale and can usually be replaced by the more interpretable *rmse*.
Root mean square error	*rmse*	0* to ∞	Square root of the mean of the squared errors.	Useful summary statistic on size of error, encompassing both bias and variability effects, in same units as forecast quantity. Gives typical size of error.	May be less useful where errors are multiplicative (ie. proportional to size of observed rainfall) rather than additive. Can be dominated by a few large errors.

Assessment Measure	**Symbol**	**Range of values** (Best value is indicated by *)	**Description**	**Advantages**	**Disadvantages**
Root mean square error of log-rainfall	*rmsel*	0* to ∞	Square root of the mean of the squared errors in log-rainfall.	Useful summary statistic on size of error, encompassing both proportional bias and variability effects, as a factor of forecast quantity. Appropriate if errors are proportional to rainfall value being estimated.	May be less useful where errors are additive rather than multiplicative (ie. proportional to size of observed rainfall). Can be overwhelmed by a few large errors. Actual interpretation obscured by use of revised logarithm.
Root mean square factor	*rmsf*	1* to ∞	Antilog of root mean square of log of ratio of forecast to observed	Appropriate where errors are multiplicative, giving meaningful scale of error. More intuitive interpretation than Root Mean Square log error. Gives typical factor by which forecasts are incorrect, so that a range can be constructed in the form: $(fcst / rmsf, fcst \times rmsf)$	Requires refinement of definition to avoid log of zero problem. Can make comparison across different sources of forecasts difficult, with danger of errors being excluded from one and not another.
R^2 (Efficiency)	R^2	-∞ to 1*	Proportion of variance in observations accounted for by forecast.	Useful dimensionless measure of forecast performance, relative to using the sample mean of the observations as a reference forecast (which gives $R^2=0$).	
R^2 for performance measures	R^2_{rmse} R^2_{mae} R^2_{rmsel} R^2_{mael}	-∞ to 1*	Proportion of improvement in performance measure of target forecast relative to a constant reference forecast.	Useful dimensionless measures of forecast performance relative to the performance potentially achievable by a constant reference forecast.	The "constant reference forecast" depends on the performance measure considered (it is provided by the sample mean for *rmse* measures and the sample median for *mae* measures).
Correlation Coefficient	*r*	-1* to 0 to 1*	Measure of linear association between observed and forecast values.	Assesses how good the forecasts might be if modified by subtracting and multiplying by constants to be selected.	Excludes effect of bias and scaling. Does not measure actual performance of forecast, only the potential performance if the basic forecast can be adjusted.

Assessment Measure	**Symbol**	**Range of values** (Best value is indicated by *)	**Description**	**Advantages**	**Disadvantages**
Categorical Skill Scores					
Ordinary Scores					
Hit Rate (Proportion Correct)	*H*	0 to 1*	Fraction of the forecasts that are correct.	Credits correct (and penalises wrong) yes/no forecasts equally.	Credits correct (and penalises wrong) yes/no forecasts equally.
Critical Success Index (Threat Score)	*CSI*	0 to 1*	Number correct divided by number forecast and/or observed	Sensitive to both missed events and false alarms	
False Alarm Rate	*FAR*	0* to 1	Proportion of forecast events that fail to materialise. 1-*FAR*=Post Agreement		Sensitive only to occasions when an event is forecasted, not missed events. Can be improved by simply under-forecasting events, at the cost of missing events.
Probability of Detection (Hit Rate for observed 'yes')	*POD*	0 to 1*	Proportion of occasions when an event does occur that are forecasted to experience the event.	Indicates ability to correctly forecast.	Sensitive only to missed events, not false alarms. Can be increased simply by issuing more forecasts, whether right or wrong.
Probability of False Detection	*PFD*	0* to 1	Proportion of occasions when an event does NOT occur that are forecasted to experience the event.		
Bias Ratio	*B*	0 to 1* to ∞	Ratio of "yes" forecasts with "yes" observations.	B>1 indicates over-forecasting (events forecasted to occur more often than observed occur) whilst B<1 indicates under-forecasting.	
Relative Scores					
Heidke Skill Score	*HSS*	-1 to 1*	Proportion of forecasts which are correct after eliminating those that would be correct on the basis of some reference (specifically, compared to a forecast in which events are forecasted to occur at the same rate they are forecasted in the actual forecasts)	Standardised scale from 0 (no skill compared with chance) to 1 (perfect forecasts).	

Assessment Measure	**Symbol**	**Range of values** (Best value is indicated by *)	**Description**	**Advantages**	**Disadvantages**
Kuipers Skill Score (Peirce's)	*KSS*	-1 to 1*	As *HSS* but the reference is constrained to be unbiased. (Specifically, the comparison is with a forecast in which events are forecasted to occur at the same rate they are occur in the set of actual outcomes)	Standardised scale from 0 (no skill compared with chance) to 1 (perfect forecasts). Forecasting rare events on basis of their low climatological probability is not penalised.	Approaches POD when correct forecasts of no-events dominate, and thus vulnerable to hedging when forecasting rare events.
Equitable Threat Score (Gilbert Skill Score, *GSS*)	*ETS*	-1/3 to 1*	Proportion improvement over chance of the probability of success relative to probability of a threat not foreseen by chance.	Standardised scale from 0 (no skill compared with chance) to 1 (perfect forecasts)	No obvious probabilistic interpretation
Odds Ratio	θ	0 to ∞*	Compares the conditional odds of making a good forecast (a hit) to those of a bad forecast (a false detection).	Good for comparing forecast performance over different time-periods, since it is relatively unaffected by differences between periods in the rates at which events occur. The log form, ln θ , can be used to split the odds ratio into the sources of benefit (eg. from events or no-events), providing improved understanding. All contributions have the same weight (not the case for *ETS*).	May be over-sensitive in cases where small numbers of forecast occasions are analysed
Odds Ratio Benefit	*ORB* θ^{+}_{ref}	0 to ∞*	Added benefit of the forecast system relative to a reference forecast. Given by ratio of probabilities involved in the forecast to those of a reference forecast.		

Assessment Measure	**Symbol**	**Range of values** (Best value is indicated by *)	**Description**	**Advantages**	**Disadvantages**
Probability Forecasts					
Categorical					
Brier Score	*BS*	0* to 1	Mean square probability error	Enables comparison of probability forecasts	Restricted to probability forecasts of exceedence of single threshold. Value of measure is not directly interpretable.
Multiple-category Brier Score	*MCBS*	0* to 1	Integrated mean square probability error		Value of measure is not directly interpretable. Probability forecasts based on different category-sets cannot be directly compared.
Brier Skill Score	*BSS*	$-\infty$ to 1*	Proportion improvement in the Brier Score of a forecast relative to a reference forecast (eg. climatology)		
Continuous					
Continuous Brier Score	*BS*	0* to ∞	Integrated mean square probability error	Applicable to probability forecasts of continuous quantities. Overall size of error is expressed in same units as rainfall.	Application requires that categorical probability forecasts be converted to continuous form.

To ensure that error quantities involving ratios of the forecast and observed quantities are always defined for zero rainfall, it is convenient to define a modified logarithm $\log^* x$: for example the *revised logarithm* shown in Table 2.2.1. Then a *modified log-error* can be defined as

$$e_i^{L^*} = \log^*(y_i) - \log^*(\hat{y}_i).$$

From this a *modified error factor* can be defined as

$$f_i^* = \exp\{-e_i^{L^*}\},$$

which is equivalent to using $f_i = \hat{y}_i / y_i$ when both y_i and $\hat{y}_i$ are large enough, but where each is replaced by a modified value if smaller than a certain threshold. It is then possible to define a *modified proportional error* ε_i^* as

$$\varepsilon_i^* = 1 - f_i^*,$$

with the *modified percentage error* being 100 times this.

The particular definition of $\log^* x$ used in this study is

$$\log^* x = \ln x \text{ for } x \geq \alpha, \text{ or } \ln(\alpha/2) \text{ for } x < \alpha.$$

Specifically, when the rainfall quantity refers to a rainfall amount in mm then α is 0.2, and when a rain rate in mm h^{-1} it is 0.8. The value of α has been set by reference to the smallest non-zero observation from a single raingauge. This choice must be regarded as somewhat arbitrary, particular when it is applied to rainfall quantities derived from weather radar when rather smaller values of non-zero rainfall are frequent. Other ways of defining a modified logarithm are available but are not considered further here.

2.2.3 Continuous Assessment Measures

Based on the above basic quantities, Table 2.2.1(a) presents a set of Assessment Measures in the form of pooled continuous variable measures involving the summation of these quantities over the set of *n* forecasts being assessed. These include forms of *mean error (bias)*, *mean absolute error* and *mean square error*. Both additive-error and log-error forms are presented. Those involving proportional (and percentage) errors are omitted as self-evident; for example, the *relative bias* follows from the formula for bias (mean error) as $n^{-1}\sum \varepsilon_i^*$, using the (modified) proportional errors. For readers unfamiliar with the summation operation indicated by the capital Greek letter "sigma", this signifies summation of the function over the samples i=1, 2,..., n.

Probably the most commonly used of this set of Assessment Measures is the *root mean square error*, or *rmse*. Table 2.2.2 indicates that it has the advantage of giving the typical size of error in the same units as the forecast quantity, and encompasses both bias and variability effects. A disadvantage is that its magnitude can be badly influenced by a few large errors, due to the squaring of the error in the summation. This is not the case for the *mean absolute error* or *mae,* which also gives the typical size of error in the same units as the forecast. A user may prefer the amplification of larger errors that the *rmse* gives, but possibly not if this is the result of atypical conditions or data error. The *mae* is said to be more *resistant* to outliers than is *rmse*.

The *root mean square factor* or *rmsf* also deserves special mention, as one variant of it is widely used in the Met Office for rainfall forecast assessment. It is particularly relevant where

errors vary in proportion to rainfall magnitude, when the typical factor by which forecasts are correct (given by this *rmsf* statistic) is more relevant than the typical error size given by the *rmse*. The specific variant used by the Met Office (Golding, 1998) is not good as a basis for comparing different forecast methods. There is a danger of not comparing like with like if suppression of trivial and zero forecasts in the summation leads to sample forecasts being omitted for one forecast method and not the other. An alternative variant suggested in Table 2.2.1 avoids this problem. The reader is left to carefully inspect the formulae in Table 2.2.1(a) and the comments made in Table 2.2.2 for each measure.

An important set of dimensionless Assessment Measures identified in the review, are based on the R^2 *Efficiency* statistic. The standard form gives the proportion of the variability in the rainfall observations that are accounted for by the rainfall forecasts, with a value of 1 obtained for a perfect set of forecasts and a value of zero for a forecast method equal in performance to a constant-value reference forecast equal to the sample mean of the observations; negative values of R^2 are clearly possible. Whilst the standard form is based on a comparison of the *mean square error* of a forecast with a constant mean-value reference forecast, alternatives are presented for log-rainfall also in terms of the *mse,* and for the *mean absolute error* for rainfall and log-rainfall when the best constant-value used for reference is the sample median of these observed quantities.

Table 2.2.1(a) presents a *Skill Score,* or *SS,* which can be used to provide a generic way of comparing a forecast method against a reference forecast for a chosen Assessment Measure. For completeness, the *correlation coefficient* is included although it should strictly not be considered as an Assessment Measure: it ignores any need to scale and adjust for bias, only measuring the degree of linear association between forecast and observed rainfalls.

2.2.4 Categorical Skill Scores

Categorical Skill Scores are designed to assess the performance of forecasts that can be judged as right or wrong, through there being a *yes/no* outcome when a forecast is compared with the observation. The occurrence of rain or no-rain is a typical *binary event* of this kind. Forecasts of actual rainfall quantities can be considered as binary events by considering the *exceedence of a threshold value*, chosen for example to be of relevance to triggering a flood warning.

The Contingency Table

Central to the assessment of binary events is the *Contingency Table* shown in Table 2.2.1(b). This is used to enter the counts *a*, *b*, *c* and *d* of the four possible outcomes for each forecast of a set of *n* under consideration for assessment. Clearly the total number of counts $a+ b+ c+ d$ must equal *n*. The four outcomes in terms of "Event Forecast"/"Event observed" are (i) yes/yes: a *hit*, (ii) yes/no: a *false alarm*, (iii) no/yes: a *miss*, and (iv) no/no: a *correct rejection*. These terms are used in some of the names of the Contingency Skill Scores constructed to form the pooled Assessment Measures. These are summarised as formulae in Table 2.2.1(b) and they are reviewed in terms of their advantages and disadvantages in Table 2.2.2.

Ordinary Scores

Arguably the most commonly used combination of scores is the Critical Success Index (*CSI*), the False Alarm Rate (*FAR*) and the Probability of Detection (*POD*). The *CSI* is the most commonly used, and gives the proportion of events correctly forecast (the *hits*) relative to the number observed and/or forecast (the *threat*). It is therefore sensitive to both missed events and false alarms. Each score has a different purpose and they are most usefully used in combination to assess forecast performance. They may be misleading when interpreted in isolation: for example a high *POD* value can reflect a frequently issued, but wrong, forecast giving a high (bad) *FAR*. The reader is left to carefully inspect the formulae for the skill scores in Table 2.2.1 (b) and the critique of them provided in Table 2.2.2.

Relative Scores

Table 2.2.1 (b) distinguishes between the *Ordinary Scores* discussed above and *Relative Scores.* These Relative Scores are constructed in relation to a *Reference Forecast.* The choice of Reference Forecast may be chance, persistence (no change) or climatology (a long-term average calculated from observations). For example the Equitable Threat Score (*ETS*) was developed as a modification of $CSI= a/(a+b+c)$ to remove the effect of the hits arising by chance, which has the expected number $a_r= (a+b)(a+c)/n$. Thus CSI is modified to $ETS= (a-a_r)/(a-a_r+b+c)$ which can be expanded to give the formula in Table 2.2.1(b). To understand how the expected number is derived, one needs to re-interpret the Contingency Table of Table 2.2.1(b) as a table of probabilities by normalising its entries by dividing each by *n.* Thus the entry counts become the probabilities

$a/n= p(f,o)$, $b/n= p(f,\bar{o})$, $c/n= p(\bar{f},o)$ and $d/n= p(\bar{f},\bar{o})$,

which sum to 1. The notation is such that $p(f,o)$ indicates the joint probability of a hit being forecast and observed (a yes/yes event) whilst an overbar signifies a "no" event (or no-event). For example, $\bar{f}$ indicates that the forecast says than an event will not occur. Note also that the marginal probability for an event being forecasted is

$p(f)= p(f,o) + p(f,\bar{o}) =(a+b)/n,$

and for an event being observed is

$p(o)= p(f,o) + p(\bar{f},o) =(a+c)/n.$

The probability of hits due to chance is

$p(f)p(o)=(a+b)(a+c)/n^2,$

and so the number of hits is *n* times this giving a_r. Thus the Equitable Threat Score is seen to give the proportion improvement over chance of the probability of success relative to the probability of a threat not foreseen by chance.

Another important Relative Score Assessment Measure is the *Odds Ratio,* θ. The *odds* (or risk) Ω of an event is the ratio of the probability *p* of it occurring to it not occurring, *1-p,* and so $\Omega = p/1-p$. The conditional odds of making a good forecast (a hit) is denoted by $\Omega(f \mid o)$, which reads "the odds of *f* given *o*". The Odds Ratio compares the conditional odds of making a good forecast (a hit) to those of a bad forecast (a false alarm), so

$\theta = \Omega(f \mid o) / \Omega(f \mid \bar{o})$.

Noting that

$\Omega(f \mid o) = p(f,o) / p(\bar{f},o) = a/c$ and $\Omega(f \mid \bar{o}) = p(f,\bar{o}) / p(\bar{f},\bar{o}) = b/d$,

we have $\theta = ad/bc$ as in Table 2.2.1(b).

A Bayesian interpretation of the Odds Ratio may be obtained by considering the Likelihood Ratio of a correct forecast of a no-event, defined as

$$LR_1 = p(\bar{f} \mid \bar{o}) / p(\bar{f} \mid o),$$

and of an event, defined as

$$LR_2 = p(f \mid o) / p(f \mid \bar{o}).$$

Note that

$$p(\bar{f} \mid \bar{o}) = p(\bar{f}, \bar{o}) / p(\bar{o}) = d/(b+d), \qquad p(\bar{f} \mid o) = p(\bar{f}, o) / p(o) = c/(a+c),$$

so $LR_1 = d(a+c)/c(b+d)$. Also

$$p(f \mid o) = p(f, o) / p(o) = a/(a+c), \qquad p(f \mid \bar{o}) = p(f, \bar{o}) / p(\bar{o}) = b/(b+d),$$

so $LR_2 = a(b+d)/b(a+c)$. Given the definitions of the Odds Ratio and the two Likelihood Ratios, it follows that

$$\theta = LR_1 LR_2 = ad/bc.$$

The odds for a correct forecast can be written as

$$\Omega(o \mid f) = p(f, o) / p(f, \bar{o}) = p(o)p(f \mid o) / p(\bar{o})p(f \mid \bar{o}).$$

This takes the Bayesian form that the *posterior odds* $\Omega(o \mid f)$ equals the *prior odds*

$$\Omega(o) = p(o) / p(\bar{o})$$

times the *Likelihood Ratio*

$$LR_2 = p(f \mid o) / p(f \mid \bar{o}).$$

That is

$$\Omega(o \mid f) = \Omega(o) LR_2.$$

Similarly, the posterior odds for a correct forecast of a no-event is

$$\Omega(\bar{o} \mid \bar{f}) = \Omega(\bar{o}) LR_1$$

where the prior odds

$$\Omega(\bar{o}) = p(\bar{o}) / p(o).$$

It follows that the Odds Ratio is given by the product of the posterior odds for a correct forecast of an event and for a correct forecast of a no-event, so

$$\theta = \Omega(o \mid f)\Omega(\bar{o} \mid \bar{f}) = LR_1 LR_2 = ad/bc.$$

It is interesting to note, in this Bayesian interpretation, that the prior odds are determined by nature, and may change substantially between dryer and wetter years, whilst the Likelihood Ratio is under the control of the forecasting method (Göber *et al.*, 2003). Thus the Likelihood Ratio itself provides a good Assessment Measure for judging the quality of a forecast method, having factored out the effect of natural uncertainty into the prior odds. The construction of the Odds Ratio as the product of the two posterior odds has also removed the influence of nature, through cancelling out the inverse but equal influences of the prior odds for an event and for a no-event.

A comparative form of the Odds Ratio in relation to a reference forecast can be constructed as the simple ratio of the Odds Ratio for the forecast method to that of the reference forecast method. This is referred to as the *Odds Ratio Benefit* and is defined in Table 2.2.1(b) and discussed in Table 2.2.2; where the *Likelihood Ratio Benefit* is similarly defined. When the reference forecast method is based on climatology (a long-term mean of observations) or chance, its Odds Ratio will equal unity and the Odds Ratio Benefit will equate to the Odds Ratio of the forecast method of interest.

2.2.5 Probability Forecasts

The categorical forecasts discussed above have concerned the forecast of an "event" defined for a continuous rainfall value as when the value exceeds a specified threshold. For forecasts that are presented as probabilities, special Assessment Measures are needed. These are summarised in Table 2.2.1(c) and Table 2.2.2. and discussed further below.

For cases where the probabilities quoted refer to whether or not a given threshold, *x,* will be exceeded, determine the occurrences of events from the observed sample of rainfall values, (y_i , i=1,2...,n), using the event indicator variable Y_i which is equal to 1 if event $y_i \le x$ does occur, and 0 if not, and where x is the threshold value. Let $\hat{Y}_i$ denote the probability of the event $y_i \le x$ occurring, as stated in the probability forecast (ranging in value from 0 to 1). Then an Assessment Measure can be constructed called the *Brier Score*

$$BS = n^{-1} \sum (Y_i - \hat{Y}_i)^2$$

giving the *mean square probability error*, analogous to the mean square error measure.

A continuous form of this Brier Score follows in a natural way by considering the threshold x to be a continuous variable. We then have an indicator variable $Y_i(\mathrm{x})$ for the event $y_i \le x$ occurring obtained from observations and $\hat{Y}_i(\mathrm{x})$ the probability of the event as stated in the probability forecast. The Continuous Brier Score is then defined as

$$BS = n^{-1} \sum \int (Y_i(x) - \hat{Y}_i(x))^2 dx \,.$$

The probability forecasts of rainfall of concern to this project are in the form of a simple probability table. The calculation of the Continuous Brier Score from these forecasts is detailed in Appendix A.

2.2.6 Overview

It is seen from this review of Assessment Measures that a variety has been developed to judge different attributes of a forecast (eg. bias, typical error size, exceedence of a rainfall threshold) and to cope with different forms of forecasts (eg. value, probability). Thus an appropriate selection of Assessment Measures will depend on the form of forecast and the users' main interests in relation to their practical application in support of flood warning. The latter is likely to differ between an informal use of the rainfall forecast for triggering a flood warning status, its quantitative use in flood forecasting and modelling systems, and its use for monitoring the quality of the rainfall forecast products for feedback purposes.

It is clear that fundamental statistics measuring bias and variability are required, pointing to the use of forms of mean error and root mean square error. Which variant to use in terms of the definition of error and its possible transformation is less clear-cut. If error size is independent of rainfall magnitude, then the standard (additive) definition of error is appropriate and the typical size of error given by the *rmse* is appropriate. For errors

proportional to rainfall magnitude a proportional error definition is appropriate, and a statistic constructed to give a typical error factor (possibly expressed as a percentage) becomes appropriate. In this case the *rmsel* (or its antilog) or one of the forms of *rmsf* are natural choices. Since rainfall forecast error size might be found to be magnitude dependent this could argue in favour of the Assessment Measures based on proportional errors. However, this type of performance measure introduces the need for modified forms of statistics that are adjusted so as to allow treatment of zero-values of forecast and/or observed rainfalls. An assessment based on the *mae* may be preferred to *rmse*-type measures if resistance to outliers is of concern. However, if datasets are large enough to contain a reasonably large set of occasions where forecast errors are large, *rmse*-type measures may be preferred due to larger forecast errors being of greatest concern to the user. Otherwise *rmse*-type measures can be dominated by the results for just one or two forecast occasions and any comparison of forecasts would be essentially anecdotal, rather than statistical.

As a dimensionless measure of performance, with an easily understood meaning, the R^2 Efficiency statistic has merit, giving the variation in the observed rainfall (about the sample mean) accounted for by the forecast method. It thus gives the improvement over the use of the sample mean as a constant Reference Forecast on a scale of 0 to 1, with negative values indicating a forecast method worse than use of the sample mean of the observations.

The Categorical Skill Scores are widely used in meteorology, particularly to judge a forecast method's ability to forecast rain or no-rain. This application is of limited interest in flood warning where the magnitude of rain is critical. Applying the scores to an event defined as the exceedence of a rainfall threshold makes them more relevant to flood warning, particularly if the threshold is chosen in relation to a rainfall threshold that might trigger an alert warning status. But if the trigger is set too high, in relation to the number and magnitude of sample forecasts under assessment, then the statistic will be poorly defined. In only judging the performance of the forecast with reference to a rainfall threshold exceedence, the skill score is failing to assess much of the information content of the forecast, and particularly the forecast maximum. This criticism can be overcome by calculating the scores for a range of thresholds, at the expense of more scores to evaluate, or each score can be pooled into a single score across all the thresholds selected.

In choosing a subset of Ordinary Scores to use, it is evident that use of a single type can be misleading: for example, the *POD* can be maximised by always forecasting heavy rain. The *CSI*, *FAR* and *POD* are a good choice to use in combination, and the *CSI* is the most useful of these and the POD least. The Relative Scores are constructed to provide a relevant baseline (a Reference Forecast) against which the goodness of a forecast can be judged: it is clearly useful to know whether a rainfall forecast product is better than a chance or climatological forecast. Ease of understanding is relevant to the selection of which Relative Scores to use. The *ETS* is attractive in giving the proportion improvement over chance of the probability of a successful forecast relative to the probability of a threat (an event forecast and/or observed to happen) not foreseen by chance. The *Odds Ratio* (and its components LR_1 and LR_2) is arguably more useful in comparing the conditional odds of making a correct forecast to those of making a wrong one (a false detection). It has merit in factoring out the inverse but equal influences of the prior odds for an event or no-event, reflecting the natural uncertainty in the rainfall and thereby focussing on the quality of the forecasting method.

The choice of Assessment Measure for use with probability forecasts of rainfall is currently restricted to forms of the Brier Score giving the mean square probability error, analogous to the *mse* measure used to assess value forecasts. Whilst the standard *BS* applies to assessing a probability forecast for a single threshold value, a continuous form has been introduced which scores performance across all possible values and which can be calculated from forecasts presented as a simple probability table.

Whilst the above overview has identified a more useful subset of the Assessment Measures reviewed, it is likely that the final choice will depend on whether the measures are for use in an automated system for assessment as part of a longer-term review of forecasts, or a semi-automated PC system with manual data entry used for more immediate within- and post-event assessment. A smaller number of measures would appear more practical for the latter. Application of some of these measures within the Case Studies featuring in Section 5 will be used to gain experience in their use, leading to the firmer recommendations for the assessment systems presented in Section 6.

2.3 Review of assessment methodologies

2.3.1 General review

This study concerns procedures for forecast assessment that are of specific relevance to the Daily Weather Forecast, Evening Update and Heavy Rainfall Forecast products. The assessment procedures that are presently employed with these products are reviewed in some detail in Section 2.3.2. Here, a more general review of forecast assessment procedures is given which reflects present international practice by national meteorological service agencies. Here, assessment procedures are primarily developed to monitor the performance of forecasts from numerical weather prediction (NWP) models operated at global, regional and mesoscales. Further assessments are made of radar rainfall forecast products for finer resolutions in space and time and for lead times typically out to 6 hours.

Past issues of the NWP Gazette published by the Met Office provide an insight into the assessment procedures in current use for assessing NWP outputs in the UK. The Met Office compile statistics compliant with the World Meteorological Organisation's Commission for Basic Systems (CBS) that allows the performance of different NWP models to be compared across the world. Statistics such as mean error and root mean square error are calculated for specific areas on a calendar month basis for pressure temperature and wind variables; rainfall does not feature. For internal monitoring purposes the statistics calculated are much more extensive, and encompass many of those reviewed in Section 2.2; more variables are considered including rainfall. Five-day probability forecasts of rainfall obtained from ensemble runs of the European Centre for Medium range Weather Forecasting (ECMWF) model are assessed using the Brier Score for different categories of exceedence of total rainfall. A major purpose of forecast assessment is to provide feedback to steer work on improving NWP model formulation. For short-term rainfall forecasts up to 6 hours based on weather radar, the Nimrod system uses the Root Mean Square Factor (reviewed in Section 2.2) as the main assessment measure (Golding, 1998).

An indication of future directions in forecast assessment within the Met Office is provided by the paper by Göber, Wilson, Milton and Stephenson entitled "Fairplay in the verification of

operational quantitative precipitation forecasts". This has been submitted for publication in the Journal of Hydrology and presents a methodology, based on the Odds Ratio and related statistics (reviewed in Section 2.2), which will be used operationally by the Met Office for assessing NWP rainfall forecasts in the future. A major advantage is seen to be the separation of the uncertainty of the event due to natural variability and that due to the uncertainty of the forecasting model. It can therefore provide a fairer assessment of forecast system performance across different years, regions and storm events for which the uncertainty due to natural variability is likely to differ. It is clearly a statistic more aimed at the rainfall forecast system developer who is looking to judge improvements in model formulation over time. Its utility for the flood-warning officer may not be so great, since here the main interest is in the accuracy of the rainfall forecast *per se* and how this impacts on the flood-warning decision-making process.

The forecast assessment procedures used by the Met Office is in many ways representative of other well-developed Met Services across the world. Consistency in procedures is supported by the collaborative activities carried out under the auspices of the WMO. The WMO-sponsored review of performance measures by Stanski *et al*. (1995) has already been referred to in Section 2.2. Another important WMO initiative is the "Standardised Verification System (SVS) for Long-Range Forecasts (LRF)" which embraces many of the performance measures already reviewed; further details are provided via the web address: http://www.wmo.ch/web/www/DPS/SVS-for-LRF.html. Another example of good practice in forecast assessment is provided by the NOAA National Weather Service "Hydrometeorological Prediction Center" (HPC) in the USA. A visit to their web site at http://www.hpc.ncep.noaa.gov/html/hpcverif.shtml is recommended for a summary of the statistics employed and near real-time displays of the assessment products. The main performance measures employed to assess rainfall forecasts are the Threat Score (CSI) and Bias (Ratio) calculated for different rainfall thresholds, and the Brier Score for probability forecasts.

The next section takes a more detailed look at the assessment procedures currently used to assess the specific rainfall forecast products of concern here: the Daily Weather Forecast, the Evening Update and the Heavy Rainfall Warning.

2.3.2 Environment Agency/Met Office Civil Centres Procedures

2.3.2.1 Introduction

The assessment methods used by Met Office Civil Centres and the Environment Agency are reviewed here to gain an understanding of the methodologies presently used in practice. This will serve as a platform from which to recommend a rationalised methodology embracing the best elements of existing practice whilst making recommendations for improvement. Information on current practice was forthcoming from the Met Office Civil Centres at Manchester, Cardiff and London and the Agency's Southwest, Thames and Southern regions. This information is summarised below and then reviewed.

2.3.2.2 Met Office Manchester

5-day forecast assessment

The assessment of the 5-day forecast employs a 5-point scoring system. Manchester Airport is used as the point of reference to assess the weather for the whole of the Northwest region. A point is deducted for each of the following errors:

Showers forecast but none occurred, or the reverse.
Forecast of rain was too slow, or too fast, by 6-12 hours.
Forecast of rain was in error by more than 12 hours (2 points deducted).
Snow occurred but was not forecast, or the reverse.

Four other error sources associated with fog, temperature, sunshine and wind attract further point deductions.

Three pooled statistics are derived from these scores:

(i) the average accuracy, calculated by dividing the total score by the maximum possible;
(ii) the number of perfect forecasts; and
(iii) the number of forecasts with up to one error.

These are looked at together to get an overall appreciation of whether forecasts are improving or deteriorating.

Heavy rainfall warning assessment

If just one raingauge within the area of warning receives the forecast amount of rain within the time period (6 or 12 hour), then the forecast is judged successful. The judgement is done on peak rainfall, not areal rainfall.

In Northwest region a "confidence level" is also provided which is intended as an informal judgement of accuracy, and not a quantitative probability of occurrence. Forecasters are not allowed to "hedge their bets" and use a 50% confidence level. A sample of 70 forecasts incorporating confidence levels have been assessed and used to calculate the probability of an event occurring at each confidence level. This is given below:

Confidence level of the warning	Probability of the event occurring.
20%	33%
30%	36%
40%	55%
60%	80%
70%	80%
80%	87%

Note that the rainfall forecasters tend to underrate their forecasts. Flood warning officers, but not the rainfall forecasters, are given this table to support their decision-making duties.

2.3.2.3 Met Office Cardiff

Met Office Cardiff provides a simple assessment system for forecasts produced for "Southwest" (Devon and Cornwall) and Wessex. The average and maximum rainfall values of the Daily Weather Forecast are compared against appropriate Met Office raingauge observations, using gauge areal averages and maxima over the appropriate area as "ground truth". Four areas are defined as "low ground" and "high ground" over Wessex and over the Southwest, delineated by the "750 foot" contour. The main assessment criterion is the % error. A recognised problem of this performance measure is that an insignificant error in absolute terms can yield a large % error when rainfall is small. The two 12 hour forecasts are amalgamated to a 0900-0900 period making it difficult to strictly compare the two 12-hour maxima in the forecast with the daily raingauge maxima; the lack of raingauge data exacerbates the problem. Bar and line charts are used to provide a visual summary of forecast performance as monthly mean % errors for average and maximum rainfall over low and high ground. A similar system is planned for Wales, but raingauge coverage over Wales is very poor.

2.3.2.4 Southwest Region

In addition to the forecast assessments made by Met Office Cardiff for the Environment Agency's Southwest Region, the Agency carry out weekly assessments using their own telemetering raingauge networks. An example provided for the two areas of North Wessex (35 gauges) and South Wessex (18 gauges) aims to assess the "min" and "max" daily forecast for the Southwest. The max is defined as the sum of the two maximum 12 hour forecasts for high ground whilst the min is the sum of the two *average* 12 hour forecasts for low ground. The latter is an unfortunate carryover from the Met Office previously providing minimum rather than average forecasts. The assessment is presented as a simple tabular weekly summary of the daily "ground truths" of the min and max for North Wessex and South Wessex alongside the forecast min and max values for the Southwest. The forecast as a % of the ground truth in the two areas is used as a performance measure range, calculated only when the forecast or ground truth rainfall is 10mm or more. Southwest Region also note that the Met Office can double-count forecast rain in the 12 hour intervals if there is much uncertainty in timing of the rain. Assessments for Devon & Cornwall are affected by using daily raingauge accumulations for an 0600-0600 day rather than 0900-0900.

2.3.2.5 Met Office London assessment for Thames & Southern Regions

The assessment of the Daily Weather Forecast employs a weekly proforma containing "Actual" and Forecast rainfalls for 7 periods (08-12, 12-18, 18-24, 00-06, 06-12, 12-24, 00-00) for three areas: Northeast, Southeast and West. The same proforma contains an assessment of the Evening Update, comparing the Actual and Forecast Maximum and Average rainfall for these three area. There is no performance measure calculated on the proforma, only space to comment. A second proforma is used to assess the Heavy Rainfall Warning, again comparing Forecast and Actual for each of the three areas: the rainfall forecasts are of the "Most likely maximum", the "Time of most likely maximum" and the "Period of rainfall". Again an opportunity to comment is given.

The Agency produce an Assessment Report on a 6 monthly basis containing an overview of the performance of the HRWs, DWFs, Evening Updates and Outlook. The performance of the DWF is categorised as follows:

Performance class	Forecast error
Very Good	within ± 10%
Good	between ± 10 and 30%
OK	between ± 30 and 50%
Poor	more than ± 50%

The percentage of rainfall forecasts in each category over the 6 month period is calculated for the time periods: Day 1 (12-24Z), Day 2(00-12Z), Day 2 (12-24Z), Day 3, Day 4, Day 5. A comment is made on whether the forecasts have improved over the previous 6 month assessment period.

The Evening Updates are similarly classified and performance commented on. A second assessment is made restricted to the occasions when rainfall exceeds 10mm. Any bias in the poor category forecasts is looked for.

In assessing the HRW, a warning issued for each area is counted as a single warning and the number issued recorded. The number (and their percentage of the total issued) of warnings assigned to one of the following 6 categories is recorded: good, over-estimated, under-estimated, issued late, timing errors, warning not meet criteria. An appraisal of the results is made.

The 6 to 10 day Outlook Forecasts is judged as accurate (or not) and whether giving "good" guidance.

The CASCADE system used by Thames Region encompasses the programs MOdaily and OSview which serve as tools for extracting information on "actual" (ground truth) rainfall relating to specified time-periods for use in assessment. (MO and OS stand for Met Observations and OutStation respectively). MOdaily calculates raingauge rainfall totals for various time-periods, and for prescribed areas gives the average (taken to be the mean) and maximum values. A calendar day is used for daily values (ie. 0000 to 0000). OSview simply displays the 15 minute rainfall data and the daily total for the calendar day.

2.3.2.6 Review and synthesis of current practice

The review of methodologies for assessment used operationally reveals that they are dominated by a small number of key features, which the Agency have found useful. These are summarised here.

First is the preference for the use of "% error". This is applied to average or maximum rainfall values for a selected area and forecast time-frame. The choice of ground-truth can influence the time-frame used and its efficacy, notably when 0900-0900 daily rainfall totals are used. The % error, as a proportional error, is undefined when the ground-truth is zero and can take large values for small values of absolute error when rainfall is slight. This problem has been suppressed by calculating the % error only when the forecast or ground-truth value is at least

a given threshold value (eg. at least 10 mm of rainfall in a 12 hour period). The % error is presented for single value forecasts and also as an averaged quantity over a "review period" (eg. monthly and as part of 6 monthly assessment reviews).

Another feature is the usefulness of "a space to comment" in simple tabulations of the forecast and ground-truth value side-by-side, in both single-value or period-average assessments. For example, comparison between review periods are made with the opportunity to comment on whether forecasts are getting better or worse.

There is also a desire to convert a quantitative assessment of error magnitude to a "performance class" expressed as a verbal ranking: very good, good, OK and poor.

2.4 Summary

Section 2 has reviewed methodologies for assessment of rainfall forecasts through a literature review and an examination of operational practices in the UK and internationally. Good existing reviews of assessment measures already exist. Consequently, this study has focussed on constructing concise tabular summaries of the measures: as formulae and as a critique of their advantages and disadvantages. While all measures are potentially useful, as they judge different aspects of forecast performance, a more useful subset has been identified for the present purposes. The experience of using some of the measures in the Case Study analyses in Section 5 will be used to finalise the selection in the concluding Section 6. The general review of operational practice in forecast assessment - carried out by national Met Services worldwide, supported by the co-ordinating activities of the WMO - and the more detailed review of UK practice, provides a useful appreciation of the state-of-the-art at an operational level from which to build in this Study.

3. Review of Daily Weather Forecasts, Evening Updates and Heavy Rainfall Warnings

3.1 Introduction

The Met Office presently issue a number of different types of rainfall forecasts to the Environment Agency on a Regional basis. Although the contents, uses and implied purposes of these forecasts differ from Region to Region it is useful to group the forecast services into three general types. These formats and contents related to these general types are discussed in Section 3.2, and some suggestions are made for how the set of forecast products can be improved. An outline of the different mechanisms for delivering forecasts is given in Section 3.3: once again these presently differ between the various Regions and between the different types of forecast product. A summary of recommendations is given in Section 3.4

3.2 Review of Current Content and Format of Forecasts

3.2.1 General

3.2.1.1 Introduction

This section (Section 3.2) contains preliminary comments on the format of the 3 types of Rainfall Forecasts: Daily Weather Forecasts (DWFs), Evening Updates and Heavy Rainfall Warnings (HRWs). The formats used for each type of Rainfall Forecast are discussed in Sections 3.2.2 to 3.2.4, where examples of the latest formats are also given.

The current formats for the Anglian, Southern and Thames regions are broadly similar with the following exceptions. The 3 forecast types for Thames region have been supplied to us as text files, while those for Anglian and Southern are postscript files: however it may well be that Word versions are available since we have been given isolated examples. The DWF format for Thames Region omits the Wind Speed tabulation contained in that for Anglian and Southern. This note relates only to the forecasts for these regions, unless another region is specifically mentioned. The example forecasts given here relate to Anglian Region and will be slightly smaller in physical size than those used operationally.

General issues concerning the three forecast types are discussed below under the headings: time reference, issue times, types of forecast, snow, consistency and forecast quantities.

3.2.1.2 Time reference

The current format for Daily Weather Forecasts contains an explicit statement that “all times are local time”. If this standard is decided upon, rather than always using GMT, we recommend that similar statements should be include at the head of Evening Updates and Heavy Rainfall Warnings. In any case, now that doubt has been introduced, the time reference used should be specified in all types of forecast. In addition we recommend that the issue time for the forecast should always be given in both local time and GMT (*i.e.* separately, even if

the times are notionally the same). This would provide a prompt for users to carefully distinguish the two time-indices and will in addition help to confirm that the change in time-index has been properly accounted for in preparing the forecasts when local time is adjusted backwards or forwards with respect to GMT. The occasions when local time is adjusted will require special care from users as some of the interval lengths will then be non-standard for the 5 days leading up to the adjustment. In addition, if a strict interpretation is to be possible, there will be a requirement to know at exactly what time the change over is made: this might not be the "official" change-over time, but possibly "midnight" or "next morning".

It would obviously be easiest if all times within a forecast were always given in terms of GMT, partly to avoid problems with the change-over but mainly because most data used within the EA for generating flood forecasts will be held assuming this convention holds. For flood warning dissemination outside the EA, the standard is to use local time and care must be used in the transformation from GMT. All warnings issued or passed on by the EA could be subjected to the same well-designed procedures for interpreting GMT as local time. Although the changeover between GMT and BST happens only occasionally, the use of a single point of translation to local time would avoid the possible pitfall of using different changeover times in different parts of a combined forecast.

If local time is to be used within rainfall forecasts received by the EA, we suggest that only the notional issue time should be defined to be in local time, with the start and end of all intervals being defined as fixed increments from this time-point. This would at least avoid problems with interpreting variable-length intervals.

3.2.1.3 Issue Times

The current practice appears to be that the forecasts issued as "Evening Updates" include only the nominal issue time of 16:00, whereas DWFs contain the actual issue-time (and not the nominal time), while HRWs must contain only the actual issue-time (since there is no nominal time in this case). We recommend that the "Evening Update" forecasts should contain the actual issue time in order to facilitate any post-event follow-up investigations, since the dates on computer files are not necessarily preserved and are possibly open to doubt according to whether or not the computer on which a forecast is originated was operating on GMT or local time. We note that the apparent actual issue-times of Updates have varied from 1½ hours before the nominal issue-time to 1 hour after. Inclusion of both the nominal and actual issue times seems worthwhile, since this can help to identify the forecast to which a revision applies. In principle, the heading information in the DWFs and Evening Updates should continue to be presented in the same format but, taking into account the need to indicate both GMT and local times, would now consist of two lines, with the first in a larger font for those systems which can make use of this facility. For example the lines might say

Forecast for : <date> at <time> (GMT), <time> (local time)
Issued on : <date> at <time>

The heading information for HRWs would be similarly modified to show both GMT and local-time versions of the issue-time.

3.2.1.4 Types of forecasts

The character of the information contained in the "Evening Update" is rather different from that provided by the DWF format and there seems no logical reason for retaining a situation in

which one format is used early in the day and another later on. In particular, if the content of the Evening Update represented by the tables of probabilities is useful, then something similar should either appear within the DWF, or be issued at the same time as the DWF. Specifically, 18-hour probability forecasts would be issued at around 06:00 and 16:00, with 5-day forecasts issued once a day only (apart from revisions as necessary). It appears that no revisions to DWFs have been issued to the 3 Regions concerned since the introduction of the Evening Updates, but it is not clear if this is by agreement between the EA and the Met Office. The examples from other regions have included cases where substantial changes to the 5 day forecasts were made in later amendments beyond the time-range covered by the Evening Updates. While there is scope for including some of this information in the "comments" portion of an Evening Update, we recommend that formal amendments to the DWFs should be issued in cases where there is a substantial change to the outlook, certainly if these are beyond the period covered by the Evening Updates or if they affect the quantities (such as temperature or wind speed) not included in the Evening Updates.

3.2.1.5 Snow

The current formats for forecasts give no indication of how forecasts of snow should be treated and the examples available contain no instances of this. We are doubtful of the wisdom of including large amounts of fixed explanatory text within forecasts, but if the present text containing definitions is retained, we recommend that the fixed rubric should contain a brief statement that snow is given in terms of its water equivalent depth. We recommend that the tabular part of the forecasts should be on this basis, if it is not already. We note that forecasts for the Midlands Region distinguish forecasted snow from forecasted rainfall by appending "S" to the numerical value of the forecast amount (separately for both "typical" and "maximum" amounts) in a tabular part of the forecast: this may prove inconvenient if an attempt is made to process the tabular material automatically and hence careful thought about specific formats would be needed.

An actual snow depth can be useful and we recommend that this should be placed in the verbal amplification portion of the forecasts. It may be best that any distinction between precipitation as rainfall or snow should be made verbally.

3.2.1.6 Consistency in terminology

We believe that it will be important to ensure that there is a good degree of consistency in the terminology used between the three types of forecasts. Some care should be taken to ensure that words used in section headings and table headings are used with the same meaning across the three types of forecast. To this end, the specifications for the three types of forecast should be set out within a single document which would contain a common set of definitions.

3.2.1.7 Forecast Quantities

It is important that the quantities that are targets for the forecasts are well-defined and well-understood. We find that the attempt to define the quantity forecasted in the "Amt" column of the DWFs is nebulous at best. The reasons for discussing this point here, rather than under the DWF heading are, firstly, that it raises the general issue that a well-defined forecast quantity should be such that a matching quantity should be calculable from observed data in some

well-defined way and, secondly, that it is necessary, as discussed in Section 3.2.1.6 to have a set of forecast quantities that are consistently defined across the different types of forecast.

The use of probability-based forecasts within the Heavy Rainfall Warnings allows a wider range of target quantities to be considered as useful quantities than is reasonable for a single-valued forecast. While the Daily Weather Forecasts contain some limited information about uncertainty, this is so poorly defined at present that the forecasts need to be treated as non-probabilistic: in addition we think that this is how they will be interpreted by EA users. In the context of probability forecasts, and sticking to amounts rather than rates, it would be possible and reasonable to ask a forecaster to produce probability tables for a "randomly selected (or typical) site", for the average amount across all sites and for the maximum across all sites. The forecast verification procedures can deal reasonably with all of these because they are judging how well the probability forecasts are doing. Unfortunately, non-probabilistic forecasts for a typical site value cannot be treated in the same way. Notionally the procedure would be to calculate an overall measure of forecast performance comparing each of the individual raingauge values (or radar pixel values) with the single forecast value. It is clear that the forecasts which are judged best would depend very much on the particular criterion used to form the performance measure: thus using the mean absolute error would favour forecasts close to the median across sites, while using the mean square error would favour forecasts close to the mean across sites. This leads to an absurdity in the overall verification procedure, since the target for a forecast quantity should not be determined by the assessment procedure that one happens to choose, but rather by the purpose for which the forecast is to be used. The assessment procedures likely to be employed are not particularly closely related to the potential uses of the forecasts, which is what would be needed in order to justify allowing them to determine the targets for the forecasts in this second-hand way. It is therefore important to specify, for each quantity in the forecast, a definite single value that this the target of the forecast. More particularly, what is needed is a rule for calculating the target value from a set of raingauge and/or radar values representing the observed outcome

We recommend that the targets for non-probabilistic, single-valued forecasts should be the average rainfall across the area and the maximum rainfall across the area. If an indication of the range of rainfall amounts likely to be experienced across an area is required then the forecast should be expressed as a range, not as a single value, and it would then be possible to devise an appropriate assessment procedure for this type of forecast.

3.2.2 Daily Weather Forecasts

The current format for DWFs is presently used for Regions which have previously been divided in 3 sub-areas for earlier formats of forecasts. The arrangement on the page is convenient when up to 3 sub-areas are used, but will need to be rethought when more are required. A repetition of the 3-area by 2-day tableau would seem to be the best available option (with minor adjustments to the number of sub-areas in each tableau for overall visual balance).

The expression of "confidence" is poor, since it fails to indicate the range within which rainfall is expected to lie with the given probability. The single "confidence" code is used for both the typical and areal-maximum values which is convenient for the format of the table, but restricts the improvements that can be sought without radically affecting the structure of the table. A possible suggestion is to define confidence ranges on the basis of a given fixed percentage (*e.g.* 75%) of outcomes lying within certain bounds based on the value quoted, where the width of the bounds varies with the confidence expressed: for example,

L = value ± the larger of 4*mm* and 100% of the value quoted ,

M = value ± the larger of 2*mm* and 50% of the value quoted ,

H = value ± the larger of 1*mm* and 25% of the value quoted .

This suggestion includes a minor variation on the straightforward percentage-error procedure which is necessary in order to deal sensibly with forecasts of zero. The same classes and bounds would be used for the 6-hour and 12-hour time-periods of the forecast. Alternative suggestions are:

(i) retain the confidence column and replace the single-valued forecasts by ranges (*e.g.* 3-6) with the interpretation that a certain percentage of outcomes will lie within the range: here both the range and percentage would be under the forecaster's control, but the values available for the percentage should be greater than 50%;

(ii) as in (i) but remove the confidence column and have the interpretation that a fixed certain percentage of outcomes (*e.g.* 75%) will lie within the range quoted;

(iii) remove the confidence column and give 3 values for the forecast quantity — a "best forecast" and lower and upper bounds for a range — where again the range would have a confidence-interpretation as in (ii).

If possible, the indication of uncertainty should be extended to include the 24-hour forecasts for days 3-5. This portion of the forecast might be improved by including minimum temperature in addition to the maximum: however the usefulness of temperature information in this form needs to be rethought, because of the altitude-dependence of relevant temperatures. Whilst forecasts for temperature are strictly outside the scope of the present project, we note that the DWFs for NE Region include forecasts for the Freezing Level which is a quantity that has a bearing on the spatial extent of precipitation falling as snow.

The expression of uncertainty for rainfall should be entirely separated from that for Lightning Risk: it seems apparent that an attempt has been made to deploy the same "risk categories" of "more than 60%, 30-60% and less than 30%" for both rainfall and lightning which has resulted in the categories for lightning surely being unsatisfactory, in that Risk of Lightning will be most commonly less than 1% for much of the year.

Discussion of the "uncertainty" and "confidence" content of the DWFs with Met Office and Environment Agency staff across a number Regions has revealed a wide range of

interpretations being placed on such information. In many instances the entries made on the forecasts are derived, more or less directly, from uncertainty assessments made of the underlying model runs. This type of assessment essentially relates to whether the weather patterns being modelled will turn out to be broadly correct in some country-wide sense and thus it would have little to do with the uncertainty of rainfall amounts over relatively small areas.

Consideration should be given to including, in the DWFs, information about immediately past rainfall, covering the four 6-hour periods leading up to the first forecast period. While this may simply duplicate information held in EA offices responsible for river floods, it is evident that the DWFs can be a useful tool for other parts of the EA without immediate access to this extra information. In addition, there are well-known difficulties in using ordinary raingauges during possible snowfall conditions, and then even a model-based assessment of precipitation can be the best available information.

5 – Day Forecast

Tel: 0845 300 0300 www.metoffice.com

Environment Agency Anglia Region (Ref: MO34)

Forecast Issued on Monday, 14 October 2002 at 06:34

Forecast for days 1 - 2 (all times are local time)

General Situation: The small area of low pressure that brought rain to the region yesterday and during the night, is currently just to the east of London. It will continue to move away eastwards this morning with a few showers following. Another area of rain, currently developing in the Bay of Biscay, will edge northeastwards overnight, bringing the threat of more rain to the region on Tuesday. There is some doubt as to the extent and intensity of rain on this system for Tuesday.

		Northern Area			Central Area			Eastern Area		
Day and Date		Amt mm	Conf	Max mm	Amt mm	Conf	Max mm	Amt mm	Conf	Max mm
Monday 14/10/02	0600-1200	3	M	6	2	M	5	3	M	6
	1200-1800	1	M	3	1	M	3	1	M	3
	1800-2400	0.5	M	2	0.5	M	2	0.5	M	2
Tuesday 15/10/02	0001-0600	0.5	L	2	1	L	4	1	L	4
	0600-1200	0.5	L	2	3	L	6	3	L	6
	1200-2400	3	L	6	4	L	10	5	L	12

Notes:

Amt:- *A typical value of measured rainfall over the Area during the period.*

Conf: *A measure of the likelihood of this value being achieved anywhere in the Area in this time period. Guidelines H=more than 60% M= 30-60% L= less than 30%*

Max: *An indication of the most likely maximum rainfall at any one location in this time period.* ***This is not an extreme value.***

Amplification: The heavy rain that has affected much of Wales and the West country overnight will edge eastwards through today. However, the rain should become increasingly fragmented and lighter with time. Many places should become dry through this afternoon. Confidence in the detail for the next batch of rain, due for the latter part of tonight, is only low. It would seem that outh and east of your region would be most at risk.

	Temperature		Lightning Risk
Day and Date	Night Min	Day Max	Low/Med/High
Monday 14/10/02	8	13	LOW
Tuesday 15/10/02	5	11	LOW

Notes:

Lightning Risk: *A measure of the likelihood of thunderstorms in the region during the period. Guidelines High=more than 60% Med= 30-60% Low= less than 30%.*

Forecast for the rest of Monday: Rather cloudy through this morning with outbreaks of rain, the rain should become more fragmented and lighter with time with a few showers and sunny intervals for this afternoon. This evening should be largely dry. Wind, west to northwesterly brisk, becoming light later.

Forecast for Tuesday 15/10/02: Rain seems likely to spread from the south during the early hours of Tuesday morning.. The rain is expected to gradually clear away later in the day. Wind, mainly north to northeasterly light increasing fresh, perhaps even strong later.

5 – Day Forecast

Tel: 0845 300 0300 www.metoffice.com

Environment Agency Anglia Region (Ref: MO34)

Forecast Issued on Monday, 14 October 2002 at 06:34

Wind Speed – 16 pt compass Beaufort Scale (Inshore waters – out to 6 nautical miles)

Day/date	Humber-Hunstanton	Hunstanton - Gt Yarmouth	Gt Yarmouth -Thames
Monday 14/10/02	N TO NE F4-6 EASING F2-3	N TO NE F4-6 EASING F2-3	N TO NW F4-6 EASING NE F2-3
Tuesday 15/10/02	E TO NE F2-3 INCREASING F4-6	E TO NE F2-3 INCREASING F4-6	E TO NE F2-3 INCREASING F4-6
Wednesday 16/10/02	NE F4-6 EASING F3 FOR A TIME	NE F4-6 EASING F3 FOR A TIME	NE F4-6 EASING F3 FOR A TIME, PERHAPS F7 LATER

Forecast for days 3-5

Forecast for 0001 to 2400 Wednesday 16/10/02:

Rather cloudy with chance of further rain, heavy at times, spreading from the south. Windy later.			
Typical rainfall mm	**Most likely maximum rainfall mm**	**Max temp °C**	**Land Wind Dir/mph**
7	12	11	NE/10-20

Forecast for 0001 to 2400 Thursday 17/10/02:

Further outbreaks of rain seem likely, heavy at times. Becoming drier and brighter during the afternoon. Windy.			
Typical rainfall mm	**Most likely maximum rainfall mm**	**Max temp °C**	**Land Wind Dir/mph**
8	15	11	N/25-35

Forecast for 0001 to 2400 Friday 18/10/02:

Mainly dry and bright with sunny spells.			
Typical rainfall mm	**Most likely maximum rainfall mm**	**Max temp °C**	**Land Wind Dir/mph**
0.5	4	10	NW/13

Outlook for Saturday 19/10/02 to Wednesday 23/10/02: Unsettled with periods of wet weather. Windy at times too. Moderate to low confidence.

Forecaster: Roger White

From Met Office London. If you have any queries regarding this forecast please phone 020 7204 7451 to speak with the duty forecaster. All times are local time.

3.2.3 Evening Updates

We think that the present format for this type of forecast makes it difficult to identify the items of most interest, which are presumably the rainfall amounts (and rates), rather than the probability table. We therefore recommend removing the need to search for these figures by introducing a new heading table and that this should be styled similarly to that at the start of the DWFs. In particular, this would deal with 6-hour periods rather than the present 18-hours, and might be extended with an additional 12-hour period so that the forecast would extend to midnight on the following day to give essentially the same termination point as the 1-2 day forecast in the DWF. It would also provide information about "typical" rainfall as well as the spatial maximum. In addition we think that information about temperature should be included and thus the tabular information at the start of the Evening Update would be rather similar to that on the first page of the DWF. Possibly the style of the verbal information should be kept the same as on the first page of the DWF, but there may be scope for replacing the "General Situation" and "Amplification" sections of the DWF by the less formal "Comments" section of the existing Evening Update.

We are doubtful of the usefulness of the probability table content of the present Evening Updates, although it does have the potential for conveying in reasonable detail the forecaster's uncertainty about future rainfall. It does also contain forecasts for instantaneous rainfall rates which is not contained in the current DWF format. While these types of information are certainly relevant in the context where HRWs are issued, it remains to be seen whether EA users find them either easy to use or useful when contained in Evening Updates. If not, then they should be removed from the Evening Updates. If they are useful, then there must be scope for including the same types of information in DWFs.

There are a number of problems with the formats of the probability tables which are essentially the same as ones arising in the HRW format: thus these comments are not repeated here.

EA Evening Update

Tel: 0845 300 0300 www.metoffice.com

Environment Agency Anglia Region (Ref: MO34)

Forecast Issued on Monday, 14 October 2002 at 1600

Evening Update.

Rainfall Amounts - Guidance for Period 1800 14/10/02 to 1200 15/10/02

Probability Of Rainfall Amounts %

Rainfall Amounts (mm)	Northern	Central	Eastern
>0-10	90	100	100
10-20	10	30	20
20-40	0	0	0
40-60			
60-80			
80-100			
100+			
Most likely maximum rainfall	4	8	8

Probability of at least this rainfall rate %

Rainfall rates (mm/hr)	Northern	Central	Eastern
4	30	40	40
10	0	10	10
20		0	0
50			
Most likely maximum rainfall rate	2	5	5

Comments: Another area of heavy rain will reach the south coast of England in the early hours of Tuesday and push steadily north. At the moment the heaviest rain in the EA Anglia region looks set to fall after 1200 tomorrow hence totals above represent the values expected from the commencement of the rain (about 0600 in southern areas & 0900 in northern areas) and 1200 only. It also follows that the figures refer more to the south of each area. A warning will probably have to be issued for tomorrow's rainfall event nearer the time.

Where probability figures indicate a need for more detailed information, please contact Met Office London forecasters on 020 7204 7451. All times are local time.

Forecaster: Andy Cutcher.

3.2.4 Heavy Rainfall Warnings

We think that the present formats for this type of forecast again makes it difficult to identify the items of most interest, which are presumably the rainfall amounts (and rates) and timings, rather than the probability table. We therefore recommend removing the need to search for these figures by introducing a new heading table which would contain these values, plus a brief qualitative indication of expected conditions: widespread/local, rain/snow, high-intensity/low-intensity, short-duration/long-duration.

The table headings, and row-information and column-headings associated with the probability forecasts are very poor and introduce confusion. In particular, for the table relating to rainfall amounts the table heading of "Probability of Rainfall Amounts" and the columns consisting of ranges (*e.g.* 10-20 mm) indicates that the second column should contain the probabilities that the rainfall amount will fall within the given range. Instead it is evident from all the examples of HRWs across the 3 regions that the column actually represents a set of exceedence probabilities: that is, the probability that the rainfall will be greater than the lower bound of the range given. Thus a heading and a table style similar to that used for rainfall rates would be better: that is, "Probability of at least this Rainfall Amount" and the column of rainfall amounts should contain just the single value (but possibly given in the form "> 0", "≥ 10", "≥ 20" if that is what is meant). The column heading consisting of "Probability of this amount at any location in the area" serves only to mislead and confuse. It seems to imply that the probabilities might apply to rainfall for a typical site, whereas it seems (but only from the fact that the row labelled "most likely point maximum" contains values corresponding to the median of the distribution) that the quantity concerned is actually the spatial maximum rainfall. A similar point applies to the similar heading in the table for rainfall rates.

The headings for the probability tables should make it clear exactly what quantities the tables refer to. Thus a sensible table heading might be "Maximum Rainfall in Area", with a possible sub-heading of "Probability of at least this Maximum Rainfall", and with a column heading of "Probability that max rainfall exceeds the given value (%)". We suggest that "amount" should be avoided in this context (including in the heading for the first column) because of its association with the "typical" value in the DWFs. Similarly, a sensible table-heading for rainfall rates might be "Maximum Rainfall Rate in Area", with a sub-heading of "Probability of at least this Rainfall Rate", and with a column heading of "Probability that max rate exceeds the given value (%)".

We suggest that the set of categories for rainfall rates should be extended to include an additional higher value of 80 mm/hr because the examples available contain several cases where quite large probabilities have been given for rainfall rates of over 50 mm/hr. The additional category would allow forecasters to express their uncertainty better in cases where such large rates are thought possible.

A decision is required as to whether a third probability table should be given for rainfall at a "typical site". This would potentially help to distinguish frontal from convective events, but the distinction between widespread and localised rain might be better made in the verbal portion of the forecast. Another alternative would be to introduce an additional information field such as "spatial coverage of rainfall (%)".

We note an instance in Southern region (3 August 2002 14:15, Kent) where a warning for two successive events was contained in a single HRW: in fact two sets of timings for start and end of events were given, but only single sets of forecasts for rainfall amounts and rates for both events together. This suggests the possible need to revise the tabular format to have columns for "Event 1" and "Event 2" so that successive events can be dealt with effectively; an alternative is to have a predefined strategy for such cases which would specify the issuance of separate warnings for each event

The examples for Northeast Region indicate that the overall objectives for this region are substantially different from those of other regions, at least when concerning the target lead-times at which warnings are generated. Lead-times for forecasts issued for the Northeast Region have often 24 hours or more, with an instance of 38 hours occurring. Forecasts for other regions have not exceeded a 24 hour lead-time. (Lead-time here is counted as being the time from the issue of the forecast to the time for the start of the event stated in the forecast.) The long lead-time for forecasts in Northeast Region has led to situations where the time-period covered by a previous HRW has not been reached when an additional HRW for a later period has been issued. The HRW strategy for this region has been to issue a HRW for the new period without providing updated information about any intervening event.

Consideration should be given to changing the way in which "uncertainty of timings" is expressed: we think that in practice the uncertainty about the start of an event will be smaller than that for the end (since the start may already have occurred). It would be simplest to give the uncertainty directly for each of the start and end of event, and one possibility would be to use a combined version of the existing formats: for example "14/1200 +/- 3hrs" (where "14/1200" is interpreted as 12:00 on the 14th). We note that the uncertainty of the time of the maximum rate is given in a different format and we think that the same format should be used in each instance. We suggest that giving the uncertainty as a range (*e.g.* 12:00-15:00) would be most familiar. We note that the examples of the existing format have some inconsistencies in the way that the time within the day is represented.

The examples we have been given reveal that contradictory information can and has been given in the probability tables for rainfall amounts and rates. We recommend that, if possible, automatic checking procedures should be implemented at the Met Office to prevent this happening. Logically the probabilities of a zero rainfall amount and of a zero rainfall rate should be the same: while these quantities are not both directly calculable from the probability tables, related quantities are and they must bear a certain mathematical relationship to each other. As an example, consider the HRW for the Kent area of EA Southern Region for 3 August 2002 (issued at 14:15). This gives the probability of a rainfall amount greater than zero as 70% (this is quite likely to be a typing error for 90%), which means that the probability of zero rainfall is 30%. However the probability given for a rainfall rate of over 4 mm/hr is 90% which would mean a probability of only 10% for a rate of less than 4mm/hr which means that the probability of zero rainfall cannot be more than 10%. Thus the "amount" and "rate" tables contradict each other.

In addition to the above problem of consistency in the probabilities, the HRWs for Southern Region contain instances where the start- and end-times of events have clearly been entered incorrectly, for example an event starting in the evening and ending in the morning of the same day.

The HRWs issued by London Weather Centre, unlike those for other Regions, do not contain an element giving a number for the warning within an overall sequence. This appears to be a useful feature both for operational purposes and for subsequent investigations. Some special consideration is needed of the specification of sequence-numbers in the situation, as here, where separate warnings are issued for each of several sub-areas. In addition, consideration should be given to including, on each of the sub-area warnings, an indication of whether or not warnings for the other sub-areas have been or are about to be issued.

EA - Heavy Rain Warning

Tel: 0845 300 0300 www.metoffice.com

EA (Ref: MO34)

Warning Issued on Sunday, 13 October 2002 at 10:51

Heavy Rain Warning for EA Anglia - Northern Area

From the Met Office London. Telephone 020 7204 7251

Probability of Rainfall Amounts

Rainfall amounts (mm)	Probability of this amount at any location in the area (%)
> 0 - 10	100
10 - 20	80
20 - 40	50
40 - 60	20
60 - 80	5
80 - 100	
100 +	
Most likely point maximum	30 mm
Start of event	13/1200
End of event	14/1200
Uncertainty of timings	+/- 3 hrs

Probability of at least this Rainfall Rate

Rainfall rates (mm / hr)	Probability of this rate at any location in the area (%)
4	90
10	50
20	10
50	
Most likely point max rate	10 mm / hr
Time of max rate	15-03 hrs

Amplification

Bands of heavy rain gradually extending north this afternoon and evening, becoming slow moving overnight across some areas, chiefly north of the Wash

3.2.5 Future Developments

3.2.5.1 Introduction

This note has concentrated on the existing formats being used now for some EA Regions and being suggested as the basis of a standard format for national use. In this final section we note some issues that may arise in the longer term.

3.2.5.2 Archiving & Performance Monitoring

One major aspect of this project is the question of performance monitoring of the forecasts. The implementation of reasonably automatic procedures for doing this requires that the forecasts should be available in a reasonably convenient form for the assessment procedures. We suggest that consideration be given to creating a separate text-version (in a well-defined format) of at least the quantitative content of the forecasts, in addition to the forecasts as actually issued, since these are not suitable for automatic procedures. The specific reason for creating and archiving text versions of the forecasts would be for the automatic forecast performance monitoring procedures. While text versions of the forecasts may be of use to the EA as a way of accessing forecast rainfall for use within an automatic forecasting system, we envisage that this role should be played by a separate forecast product providing more detailed spatial and temporal information and not meant for direct human interpretation. The text version of the current forecasts would implicitly define the quantitative and categorical information available for forecast performance monitoring, and hence would have to chosen with care.

The need to have forecasts available for ready use within an automatic forecast-assessment procedure leads to a corresponding requirement for the archived forecasts to be held in a suitably restricted format. Of course, some flexibility can be built into the procedures for accessing data within the assessment programs, but even then there are limits to what can be interpreted in a reliable way. While forecasts are often generated and sent out to regions by forecasters who use computer-tools for this purpose, it seems that these tools do not necessarily perform any checking of entries made into them, but rather just pass the data on in a simplistic way. It would be worthwhile improving the capabilities of these tools in order to ensure a consistent format of the files to be archived and/or distributed: part of this task would involve identifying exactly what the forecasts should contain, or should be allowed to contain. Some examples of potential problems with the contents of forecasts found, either within the present study or in other experiences with operational forecasts, are as follows.

(a) A field which nominally contains a single value contains what is evidently meant to be a range, for example “10-15”.

(b) A field which usually contains just a numerical value, contains in addition extra characters indicating the units, for example “mm/hr”.

(c) A field which usually contains a numerical value which is a “whole number” contains a number in decimal format, for example “0.5”. In this case the answer may just be to ensure that the assessment procedures expect to receive such data- fields.

(d) A field which usually contains a numerical value contains some erroneous characters such as “O” instead of “0”.

(e) Data-fields within a forecast may provide inconsistent information, such as in the case of the probability-forecast content of the Heavy Rainfall Warnings for which some problems noted have been discussed at the end of Section 3.2.4.

(f) Fields which nominally contain times of day or, more generally, date and time, may have varying formats for their contents. For example, a time may be expressed as "12:00", "1200", "12" or "12 hrs". In addition, there have been cases where unusual values such as "0001" or "2359" have been used where there were evidently problems in knowing what are the valid ways of expressing values at the beginning and end of a range.

Problems such as these are relatively unimportant when a forecast is used only for visual inspection, but could have serious consequences for automatic procedures.

3.2.5.3 Restructuring of Forecasts

We think that further development of the Heavy Rainfall Warnings should be considered, in particular to provide more useful information about the likely timing of rainfall within an event. One suggestion is that a time-profile of rainfall amount should be provided using 3-hour intervals which would be at fixed times of the day. There are indications in the example HRWs that forecasters have sometimes felt the need to be able to put over the idea that more than one period of rainfall is expected, and this would be one way of allowing this information to be expressed in a quantitative way. We think that such a radical change to the information contained in HRWs would need to be subject to initial local trials to assess the feasibility of the requirement.

The development of the three distinct rainfall-forecast products issued to the Environment Agency by the London Weather Centre has perhaps led to the rather unfortunate tendency for these to be considered as separate services specifically required by the Environment Agency. In practice, the Met Office products need to be considered as part of a single overall service providing relevant information about the rainfall likely to be experienced in the immediate and more distant future. This affects not only the way in which the services should be formatted and structured, but also the way in which the products can be interpreted by Environment Agency staff. For example, the fact that a Heavy Rainfall warning has *not* been issued implicitly provides an upper bound to amount of rainfall that is likely to occur and this which could be used to complement or revise any information contained in the last regular forecast. However this is made difficult by the use of incommensurate systems of defining time-periods in the forecasts, by a lack of knowledge about the timeliness and accuracy achieved by Heavy Rainfall Warnings and by a lack of information about when a Met Office forecaster has checked whether or not to issue a warning.

At present the HRW service is defined in relation to certain thresholds of rainfall amounts or rainfall rates which are effectively targets for occasions when HRWs should be issued. It seems that the contents of the HRWs issued by London Weather Centre have become rather disconnected from these basic trigger events, in that the HRWs do not explicitly say why, or against what criterion the Warning has been issued. It is arguable that this disconnection is unimportant and of concern only to the question of measuring the performance of forecasts rather than to the more important task of providing useful real-time information to Environment Agency staff. However, as discussed in Section 3.2.4, the existing HRW formats lack the ability to provide the sort of temporal resolution that the Met Office forecasters are attempting to supply and that would be useful to the Environment Agency. This partly arises from the structure of the HRWs emphasizing, firstly, the identification of a "rainfall

event" and, secondly, the forecasting of the total amount over the event period. In contrast, the specification of when HRWs should be issued is usually in terms of running totals of rainfall amounts over a selection of interval-lengths. Logically, this should lead to the Warnings containing, for each such interval-length, a list of those times-points for which the running-total ending-then would exceed (or nearly exceed) the given thresholds. The specification of the criteria for when Warnings should be issued should be made in conjunction with deciding on the contents of the Warnings. This process should make a strong distinction between criteria based on the average rainfall across the sub-areas and those based on the worst cases over the sub-area.

It seems that the present HRW service does not usually provide routinely for updates of forecasts once an initial Warning has been issued, nor for a formal down-grading of the warning situation. It may certainly be that part of this is covered in telephone conversations between Met Office and Environment Agency staff. Nonetheless it seems an obvious missing element to the service that formal updates are not usually provided, if only so that the Environment Agency receives confirmation that the warning-situation is still on-going. An update every 6 hours may be about the right frequency if conditions are not rapidly changing. If the Warnings are restructured so that there is less emphasis on forecasting for a whole-event at once, there would be no clear "end of event" predicted in forecasts already received and thus there would be a need to provide a way of bringing a "warning-condition" to a close. At present, the way in which the end of a warning period is to be interpreted seems unclear: it is not obvious if the warning-period ends at some point at which the forecaster no longer has sufficient certainty to issue a prediction, or low- or zero-rainfall is being forecasted beyond the stated time. The removal of the need to deal with separately-identified rainfall events within the HRW forecasts would overcome some of the problems of dealing with closely-following events that were discussed in Section 3.2.4: each forecast would cover a possibly-variable overall time-period, but extending out to at least the time-point for which the need for a warning-condition can be predicted with reasonable confidence.

3.3 Delivery Methods

At present the rainfall forecasts provided by the Met Office to the Environment Agency are delivered by a variety of mechanisms. These differ between the Agency's Regions and between the different types of forecasts. In some instances, the forecasts provided to a given Region by the Met Office are sent via a number of mechanisms and to a number of target addresses. Once received by the Agency's Regions, the forecasts may be subject to forwarding-on within the Region, either by forwarding to particular post-holders or by posting for more general availability on an intranet service or other computer notice-board or shared-folder arrangement. Again the arrangements within a Region for onward delivery may differ between the different types of forecasts.

The visual formats and contents of forecasts differ between the Agency's Regions and have been discussed in Section 3.2. When considering how forecasts are delivered, the word "format" is also used in relation to the type of computer file used as part of the delivery process. Three different file-formats are used in connection with the forecasts sent to the Agency's Regions, with a fourth being used for the "National" forecasts which are not considered elsewhere in this report. These file-formats are plain text (ASCII text), Microsoft Word format, and Postscript for the Regional forecasts, while the National forecast uses HTML. There are obvious advantages in using file-formats other than plain text, as these allow visual aids to be used to separate the different portions of the forecast and to highlight the most important parts. However, plain text has the advantage that this is at least usable where distribution within a Region relies on a computer system which doesn't have a GUI-based user-interface. Further, plain text has the important advantage that files in this format can be readily used within computer-based systems for making further numerical use of the forecasts, as opposed to just displaying the forecasts. This relies on certain parts of the forecast being provided in a prescribed style so that the relevant information can be extracted. There can be real-time uses for this type of information: for example in the RFFS system used by North East Region, where the Daily Weather Forecast is used via automatic procedures to provide information on future rainfall to the flow forecasting system. (However, future advances in methodology are likely to lead to rainfall information specifically for use by river flow forecasting systems being provided at a much more detailed spatial scale than would be appropriate to the context in which the forecasts dealt with here are used.) An additional use for plain text forecasts is likely to be in automatic procedures for monitoring the performance of the rainfall forecasts.

The mechanisms for delivery of the rainfall forecasts from the Met Office include FTP, e-mail and Fax. Each of these is in use for delivery of the Daily Weather Forecasts and Evening Updates (which usually share a common delivery mechanism within a Region), and for the Heavy Rainfall Warnings. There is not a strong relationship between the file-format and the method of delivery, although in most cases plain text forecasts are sent by FTP and Word-format files are sent by e-mail. The Daily Weather Forecasts and Evening Updates are usually sent by either FTP or e-mail (or both), although at least one Region receives a copy by Fax (in addition to a copy via e-mail). Heavy Rainfall Warnings are sent by Fax to most Regions, although one region receives these by FTP (only) and another Region receives copies by both Fax and e-mail (for archiving).

Where FTP is used as the mechanism of delivery, there is essentially just the one point of delivery within a Region with onwards transmittal being under the Agency's control by

automatic procedures. When Faxes are used for the delivery these are usually sent to the appropriate Regional Control Centre, and then retransmitted as Faxes to area offices and to flood duty staff. In the case of e-mail delivery, it would be relatively easy to arrange for delivery to multiple addresses directly from the Met Office. However, it seems that this facility is only used in one Region and even then the messages are forwarded-on to Flood duty officers in much the same way as in other Regions. Maintenance of the lists of those who should finally receive the forecasts and warnings seems best done within the Agency's Regions.

3.4 Summary of Recommendations

The following is a summary of the main recommendations relating to the format and content of the Regional rainfall forecasts services provided to the Environment Agency.

(1) The contents and format of the different types of forecasts should be considered and specified jointly, so that consistent definitions and terminology are used.

(2) All time-periods within the forecasts should be specified directly on the GMT scale. Actual issue-times for forecasts should appear explicitly within each forecast and these issue-times should be given on both local and GMT time-scales. It is important that the time-scale being used within the forecast should be explicitly stated, for each of the different types of forecasts. Regular forecasts might also include the nominal issue-time for identification purposes.

(3) The format and content of the present Evening Updates should be entirely replaced by adopting instead a shortened form of a revised Daily Weather Forecast where, for Evening Updates, this would be restricted in time-coverage to finish at the end of the next day.

(4) The format of the present Heavy Rainfall Warnings should be revised initially so that the important parts of the forecast become relatively more prominent, rather than being mixed in with less important details. The important parts would be the time-period of the event and the amount of rainfall.

(5) In the slightly longer-term the formats and contents of the Heavy Rainfall Warnings should be revised to reflect in a useful way the agreed criteria for when Warnings should be issued. For example, where a criterion is that a Warning should be issued if a 12 hour total rainfall at a site is expected to exceed 20mm, then the Warning should be capable of expressing this and of indicating when such a period is forecast as occurring (e.g by stating the end-points of such 12-hour intervals). Typical sets of criteria for issuance of Warnings refer to several different interval lengths, and all should be covered by treating all criteria separately within a Warning and by allowing an indication that exceedence of the threshold for that criterion is not expected.

(7) The contents of the Heavy Rainfall Warnings for regions other than those served by the London Weather Centre can be improved as an interim measure by changing the style of the verbal messages to provide more specific information about the timing and amounts of the events being forecasted. In particular, nebulous phrases such as "early tomorrow" should be avoided unless backed-up by referring to a particular hour of the day (or range of hours). The

point here is that such phrases can be ambiguous: for example does "early in the day" refer to the calendar day, daylight hours, or typical working day? Clearly, where there is any uncertainty in timing, this should be made clear in the forecast.

(8) The present use of Heavy Rainfall Warnings by the Agency's North East Region is rather different from that employed elsewhere and is probably best treated by inventing a new name for a service of this type, such as "Long-Term Warning". This should be followed by considering whether a similar service would be useful for other Regions. For most Regions the time-horizon covered by Heavy Rainfall Warnings is restricted to events starting within about the next 48 hours, while those received by North East Region seem targeted at a 2-5 day horizon. In principle, the intention of (or need for) these Long-Term Warnings should be adequately covered by the Daily Weather Forecasts but the receipt of a separate Warning may provide more flexibility for the Environment Agency in getting the information to those who do not normally receive the Daily Weather Forecasts.

(9) The principal target quantities for forecasts should be the average rainfall within an area and the maximum rainfall, where these would both be total rainfalls over prescribed periods. Where rainfall rates need to be targeted, careful consideration is needed of the relevant space and time-scales for these: we suggest that the smallest realistic scaling would be to define rates in terms of averages over 2×2 km^2 radar-pixels and over 15 minute time-periods.

Addendum

At the time of finalising this report (June 2003), certain changes to the format of operational forecasts provided by the London Weather Centre had already been made. These changes are summarised below.

Daily Weather Forecasts. Changes have been made to include forecast information for longer time-periods for the initial two days, in addition to the 6-hour and 12-hour periods in the format discussed above. Thus an 18-hour total (06:00-24:00) is given for Day 1, and a 24-hour total (00:00-24:00) for Day 2. The fields representing "Lightning Risk" have been removed.

Evening Updates. Additional forecast-values have been provided. These give "typical" values of the rainfall amount for each Area in addition to the existing set of "most likely maximum values". This makes the interpretation of the probability tables less obvious in terms of the rainfall quantity they represent. A corresponding change to the format of HRWs has not been made.

Heavy Rainfall Warnings. Provision has been made to list the agreed criteria for when Warnings should be issued, and for check marks to be given for those against which the particular Warning is made. In the text version of the Warning, this portion is not particularly easy to read. There may be problems in formatting this type of information for other Regions which have a larger number of threshold criteria. The information provided by these check boxes might be extended further to indicate directly, for each threshold, the time-periods for which the threshold will be exceeded, as suggested in this report. This would allow forecasters to warn of the occurrence of adjacent events.

4. ASSESSMENT PROCEDURE FOR USE WITH CASE STUDIES

4.1 Introduction

The problem of assessing the performance of rainfall forecasts can be divided into a number of basic issues:

(i) the questions to be answered by the assessment;
(ii) the target quantities or events;
(iii) the assessment framework;
(iv) selection of data for assessment;
(v) evaluation of the assessment.

These points are addressed below and some of them are expanded upon in later sub-sections.

(i) The questions to be answered by the assessment

An assessment procedure clearly needs to be designed on the basis of the underlying purpose of the assessment. An analysis of forecast performance may be required to address one or more of the following slightly different questions, each of which has a different implication for how the assessment procedure should best be framed.

(a) Find the typical types and sizes of error in the forecasts from a given source. A use for such an analysis might be to provide an indication of the sizes of errors to be expected in future forecasts made in similar circumstances.

(b) Compare the performance of forecasts from a given source over different periods of time. Here the aim might be to detect whether the effort put into supposedly improving forecasts has had a noticeable effect.

(c) Compare the performance of forecasts from two given sources over the same set of forecasting occasions. The aim would usually be to find which source gave better forecasts. The analysis might involve entirely separate forecasting services, or the problem might be one of testing whether a minor variant of an existing service has achieved an improvement.

(ii) The target quantities or events

In order to implement a forecast-assessment procedure there must be some clarity about what it is that is being forecasted and some way of matching the items being forecasted against a corresponding outcome that can be determined after the event. Where the target of a forecast-service is specifically to forecast the occurrence of some type of rainfall event, it is usual to characterise the occurrence or non-occurrence of the event in terms of some more quantitative measure of rainfall amount or rate. Thus the specification of the targets for quantitative forecasts is of prime importance.

A target quantity for rainfall forecasts is defined by the location and resolution of an interval in time and space and, implicitly, by its method of derivation from a notional function representing the rainfall intensity at an arbitrarily-fine resolution. The spatial part of the specification of a forecast may be, for example, rainfall at a single special location, the average over a particular area, or the maximum over an area. A number of other possibilities are available which take different approaches to the problem of defining a useful quantity to use as a target, given that rainfall may well vary substantially over a given area. The temporal

part of the specification of a forecast may be a single time-point, the average over a particular time-interval, or the maximum over a time-interval. Here the suggestion of using the maximum rainfall intensity as a target quantity can be considered as an attempt to define a useful and meaningful item within the forecast that, similarly to the spatial case, acknowledges that rainfall intensity may vary considerably over a time-period. However, the use of a maximum rainfall intensity raises important issues when attempting to define a meaningful target. If the terminal velocity of a raindrop is taken to be about 7-9 ms^{-1} (for drops of sizes 2 to 1.5mm radius), this is equivalent to a rate of 25-33 $\times 10^6$ mm h^{-1} over the very small area of arrival and during the very small time-interval in which the raindrop touches ground. The practical purpose of a forecast of maximum rainfall rate requires that there be some averaging in time (and possibly also in space) and the specifics of this should depend mainly on the use to be made of the forecast. There can be some attraction in matching the detailed specification of the maximum rate to the ground-truth available for assessing the forecasts, or to a representative ground-truth with which the forecast-users are familiar.

(iii) The assessment framework

The way in which an assessment of forecasts is implemented should be strongly influenced by the first two factors above. There is clearly a need to acquire both forecasts and values for the outcomes of the target-quantities in the forecasts. For the present study, there is a substantial interest in being able to compare the performances of different forecasts that are likely to have only relatively minor differences. However there is also an interest in simply being able to monitor the performance of forecasts, partly as a guide to the sizes of errors to be expected in future forecasts but also as a way of continuously monitoring the quality of the forecasts being received by the Environment Agency.

There are two main requirements for comparing different sources of forecasts. Firstly, essentially the same set of outcomes should be used to construct performance measures for the two forecast sources, since this means that a like-for-like comparison can be made. This requirement may rule-out certain versions of performance measures where a reduced set of forecasting-occasions is constructed based on the values being forecasted: two different forecast-sources might mean that the sets of forecasting-occasions are different and hence the performance measures would not be directly comparable. A second requirement is that the assessment of the different forecast-sources should ideally take place within a single overall procedure, rather than simply evaluating performance measures separately. It should be possible to design an overall procedure which will allow a statistical analysis to be made of whether or not there is enough information in the data-sample to determine if one forecast source is better than another. For the present project, this ambition has been implemented for only a subset of the performance measures being considered, but it should be generally achievable provided that an appropriate methodology can be developed. Section 4.3 outlines the method of comparing forecasts that has been used, and it goes on to consider possible other methods of comparison for performance measures where this cannot be applied.

The need for a forecast assessment procedure to have available simultaneous forecasts from all the forecast-sources being considered highlights an important point regarding the analysis of the Heavy Rainfall Warnings (HRWs). For simplicity, the approach to analysing HRWs here has been to use the Warnings actually issued as the basic set of occasions to be analysed and to concentrate on assessing the accuracy of the forecast-quantities within each Warning.

Clearly, different potential sources for generating equivalents to Heavy Rainfall Warnings would be likely to lead to warnings being issued for different sets of events, or at different time-points within an event. This points to the need for a different type of analysis. It is further clear that the analysis only of Warnings-issued omits consideration of cases where high rainfall occurs but no warning is actually issued. An alternative approach to the analysis of HRWs is potentially available but it has not been possible to implement it within this phase of the present project. In the simplest case, the approach would be based on considering all time-points within an overall time-period, for example at hourly time-points, and determining from observed rainfall data whether or not a Warning should have been in force at that time point. An analysis could then be made of the actual periods for which Warnings were in force. A number of alternatives are possible: for example, the initial identification might concern those ranges of time-points within which the Warnings should ideally have been issued. The practical implementation of such schemes has been partly thwarted by the fact that the Heavy Rainfall Warnings are not themselves well-designed for this type of analysis. Thus there are multiple criteria for when Warnings should be issued, but the Warnings do not indicate against which, if any, of these criteria the warning is issued. Further, the time-periods covered by a Warning do not seem to have a well-defined relationship to the times at which the notional thresholds for issuing warnings are expected to be breached. This makes the implementation of schemes such as those described here problematic and a wide-ranging exploration of possible ways of interpreting the present Warnings would be needed.

There is the additional problem that ongoing revisions to the formats and contents of the Warnings, and changes to the criteria and to the types-of-criteria used for issuing Warnings would be likely to lead to the need for a complete re-evaluation of how the assessment should be implemented. While it is important that an assessment of this general type should be made (i.e. one which allows an analysis of possible failure to issue warnings and of the timeliness of warnings), further thought is needed about the overall purpose of the assessment and, in particular about whether an assessment should be made of the joint performance of all the forecast-services that the Environment Agency receives from the Met Office: this might aim to provide an assessment of the knowledge provided by the latest Daily Weather Forecast and/or Evening Update, together with information gained by whether or not a Heavy Rainfall Warning has been issued.

(iv) Selection of data for assessment

The way in which a dataset is selected for an assessment of forecast performance can have a major impact on the interpretation of the results. For example, it is fairly common for a dataset to be chosen so that it contains a number of notable rainfall events, or at least to tend to exclude long periods of rain-free conditions: a primary objective of this selection may be to avoid analysing long-periods when both forecasts and observations are zero. It should be recognised that, if the purpose of the analysis is to provide a representative size of error for future forecasts, the results will only be correct (or be interpreted correctly) if the selection of the dataset is taken into account: i.e. the errors in the sample forecasts will only be representative of forecasts made “during rainy periods”, or of whatever the dataset is considered representative. In principal it is better to analyse a dataset covering the whole of a long period of data and to provide a result for the likely size of the error in a way that is conditional on the value forecasted, rather than averaged across all occasions: thus the result might be that the error is likely to be ±4mm if the forecast is 4mm and ±8mm if the forecast is 20mm.

The discussion above, under “assessment framework”, has already indicated that where there is a need to compare forecasts from several sources, there are major advantages in having a dataset consisting of occasions when forecasts are available from all sources.

(v) Evaluation of the assessment

One of the problems with evaluating a set of performance measures for a set of forecasts is that the raw measures of performance are usually presented without any indication of how well the numerical values are determined by the possibly very limited dataset available for the analysis. This makes the comparison of forecast performance over different periods of time somewhat problematic although, if enough time-periods are considered simultaneously it should be possible to identify what component of any apparent change is just random noise. The discussion above, under “assessment framework”, has already indicated the possibilities of extending the analyses within an assessment framework to include an internally-derived measure of the accuracy with which performance measures (and differences between performance measures) can be estimated.

One problem with some of the performance measures is that valid numerical values cannot always be determined, depending on the dataset to which they are applied. While this may be overcome in some cases by providing values for the measure in ill-determined cases, this may not always be possible. A problem here would be that, even though such a performance measure might be well-defined for a given dataset, it might not be possible to find methods for deriving a measure of accuracy of the performance measure if the measure is not well-defined for all possible datasets.

In general, measures of forecast performance have not been derived specifically to answer the question of whether there is enough evidence in a dataset to distinguish the forecast-performance of forecasts from two sources. It seems possible that new measures specifically for comparing performance can be derived.

4.2 Choice of Ground Truth

Any procedure for assessing the performance of forecasts must make use of some dataset that identifies the actual outcomes for the occasions on which the forecasts are made. In the case of rainfall, even in the best of circumstances, there are particular problems in determining the amounts of rainfall that have fallen. These are well-known and relate primarily to the sparseness of reliable raingauge networks which would otherwise be taken as good measurements of rainfall at individual locations. Other problems arise from the differing characteristics of the potential data sources, which include both raingauge and weather-radar sources, and of the possible ways in which information from these sources can be combined. Some of the properties of the primary data-sources are summarised later in this section.

An assessment of the performance of rainfall forecasts may be undertaken for several different reasons and for it may be appropriate to use a different version of ground-truth for each of these. This would usually arise from the timely availability of the required data and from the effort required for data-acquisition, quality-control and other data-processing. A clear distinction needs to be made between the analyses of forecast performance made in this report and the assessments that will be undertaken by the Met Office and the Environment Agency subsequently. The analyses in this report have been made using versions of rainfall-

ground-truth from sources that are readily available at the present time and so the ones deployed should not be taken as recommendations for later use. In particular, for reasons of time and availability, we have not included sources that merge together information from weather-radar and raingauge networks. Such sources are likely to be prime candidates for operational use. Two main reasons for undertaking operational assessments of forecast performance are identified below, noting that the appropriate ground-truth for use in such assessments may well be different:

> *(a) Routine performance monitoring.* Monitoring of the performance of rainfall forecasts may be undertaken on a routine basis, perhaps monthly, in order to provide feedback on the forecasts being made and to help to identify any problems that may arise. In addition, assessments of the forecasts received may be made immediately after noteworthy rainfall events. For these purposes the choice of rainfall ground truth will usually be determined by what data are conveniently to hand.
>
> *(b) Comparisons of forecast performance.* An important type of assessment arises where the primary aim is to test the performances of different variants of a forecasting procedure in a direct comparison. The main requirement here is for a dataset covering an extensive time-period, since it is only by using lots of data that minor differences can be revealed. Such comparative analyses are likely to be undertaken somewhat less frequently than routine assessments and in such circumstances it might be thought worthwhile putting extra effort into assembling the dataset to be used as ground truth. There are additional considerations here, in that the choice of ground-truth should not be preferential to any one of the candidate forecast procedures, but should allow the best features of each to be brought out by the assessment. For example, a variant of a forecast procedure might have been constructed so as to provide improved forecasts over high ground. There is a much greater need for the dataset used as ground-truth to match reality when used for comparing forecasts than for routine monitoring, and this needs to be taken into account when considering the resources used in constructing the dataset.

The choice of rainfall ground-truth is to some extent affected by the specific quantities that are targets for the forecasts. In the present context we exclude cases where the forecasts would be made for certain specific locations at which there happen to be raingauges. Problems may arise from the time-resolution of the target quantity. Section 4.1 has already discussed the question of rainfall rates where there is need to determine a practicable definition of what the corresponding ground-truth should be. One possibility is to use a rate calculated from 15-minute total rainfalls as the closest reasonable representation of an instantaneous rate. However, if this were adopted, this might preclude the use of some planned radar-raingauge merged products where iy is often thought that a resolution of one-hour is adequate. In those cases where the target of forecasts is specifically a rainfall accumulation, the periods chosen are typically 6 hours or more. Thus a one-hour resolution for the data used to provide the ground-truth is adequate in these cases.

Problems can also arise from the spatial resolution associated with a target quantity. Where the target is a spatial average rainfall then values from a raingauge network would certainly be a reasonable candidate for ground-truth. However, several of the existing forecasts have targets that are spatial maxima. As with the temporal case, there is a need to define point values of rainfall as averages over a small local area: otherwise unrealistically large spatial

maxima arise. It may be most natural to use one of the standard resolutions for radar data to define this local averaging when specifying the target for the forecast of spatial maximum rainfall. This raises the potential difficulty that sources of data for ground-truth may have differing resolutions. In principle, a raingauge network does not provide a good estimate of the largest rainfall within an area and so should not be used on its own to provide ground truth where the target is the spatial maximum rainfall. However, it seems from the case studies that in some cases the forecasts have been tuned to provide estimates similar to the maximum from a set of raingauges. This leaves open the possibility that the true target in these instances is actually the maximum of the values observed at a number of widely spaced locations, rather than the true spatial maximum. Here, the number of locations would correspond roughly to the number of raingauges typically operating within the telemetering network, but the forecast-target need not relate specifically to these sites, .

The properties of the basic sets of rainfall data that are available, or that are potentially available, for use in deriving ground-truth rainfalls are as follows.

(a) Daily-read raingauges

These are often taken as the major determinant of ground truth. An extensive network of such gauges exists in the UK but there can be, at best, a delay of several months arising from data processing and quality control if a large number of gauges from the national network were to be required. The Environment Agency itself operates some of the national network of daily-read raingauges registered with the Met Office, and some others, and so the Agency's access to these would be easier: however quality-control would still need to be dealt with and this is made more uncertain by having fewer gauges. The accumulation periods of daily-read gauges are 9:00 to 9:00 GMT in the vast majority of cases. Unfortunately, this time-period does not usually coincide with the intervals used in the rainfall forecasts received by the Environment Agency, even where a 24-hour period is being dealt with. Thus data from the daily-read raingauge network are not of direct use for forecast assessment. While there is a notional possibility of combining daily-read raingauges with recording raingauges to provide a representation of ground-truth that improves on both, it seems unlikely that this would be undertaken in practice. An important point is that the daily-read network of gauges has a better spatial resolution than typical operational networks of recording gauges.

(b) Recording and telemetered raingauges

These provide data at a reasonably high temporal resolution, the data consisting either of 15-minute totals or "time-of-tip". The spatial resolution available from this source is limited by the number of raingauges in the network: unfortunately, this number is usually rather small. There can be minor problems relating to tipping-bucket calibration, where the use of daily-read check gauges has sometimes been used to improve the measurements. The datasets from networks of telemetered or recording raingauges usually require extensive quality-control to overcome instrumentation problems where gauges stop operating or otherwise yield obviously incorrect results. Data received via telemetering systems are notionally available immediately they are received, although polling of outstations may only be undertaken once per day in non-flood conditions. Thus up-to-date data would be available for use as the basis of ground-truth essentially as soon as required, subject to the requirements of quality control procedures.

(c) Unadjusted weather radar data

Estimates of rainfall from raw radar sources have notionally good properties in terms of the spatial and temporal resolution. Descriptions of how weather radars operate indicate some minor problems in defining these resolutions arising from details such as the beam widths and frequency of sweeps: these are usually overlooked but may prove problematic if the targets for forecasts really were rainfall rates determined over small time and space intervals. The basic rainfall data from weather radar are subject to a range of problems that need to be corrected (persistent anomalies, anomalous propagation, attenuation effects bright band, etc.) and, even if corrections are made, estimates of rainfall amounts can be poor unless values are adjusted in relation to contemporaneous observations from raingauges. Quality-control of radar data is often required to remove obvious bad radar images, and there is then the problem of have to deal with any missing images in evaluating the final rainfall estimate.

(d) Corrected and adjusted weather radar data

The Environment Agency have available to them a number of products which provide rainfall estimates from weather radar which are quality-controlled, which implement corrections for certain of the effects described above and which make adjustments on the basis of measurements from telemetering raingauges. The principal such product is the Nimrod "quality controlled" dataset, which is primarily obtained by the Agency for the operational uses of monitoring current conditions and flood warning. This product is still undergoing development, in particular in relation to the sets of raingauges used for adjustment. The adjustment procedure does not attempt to match radar and raingauge rainfall amounts in a fine-detailed time-scale (i.e. frame-by-frame), but instead uses slowly-varying adjustment factors which are evaluated over moderately long periods of time. In particular, raingauge-rainfall for a given time-period may not be available at the time the image for that time-frame is processed. This is in contrast to the Hyrad product which adjusts radar images on a frame-by-frame basis, but which contains provision for recalculating the adjustments for past images should telemetry data arrive late. The usefulness of these sets of processed radar data for non-real-time use is something that needs to be assessed, in particular in relation to the effectiveness of the quality-control procedures: thus, for real-time use, it may be sensible to take the view that something-is-better-than-nothing and thus to allow the use of problematic data that would not be passed for other uses. In any case, it would be sensible to plan to undertake a further visual quality-control of the processed data in order to identify any problems not found by, or caused by, automatic processing procedures. Further problems may arise from the use of real-time telemetry data that have not been fully quality-controlled: use of poor raingauge data may mean that the entire dataset will need to be reprocessed for post-event analyses. There would be obvious problems in relation to this for the Environment Agency since the Nimrod processing tools would not be available in-house.

(e) Merged raingauge and radar products.

We understand that a new Nimrod product is being planned which would include a more comprehensive combination of raingauge and radar products. The documentation for this which is available suggests that this would be targeted at producing estimates of one-hour rainfall accumulations. In contrast to the existing Nimrod "quality-controlled" product where the principal aim is to provide up-to-date

rainfall estimates essentially as soon as the raw radar data are available, it seems likely that the merged product would be subject to slightly more delay arising from the acquisition of raingauge data from the Rainfall Collaboration Project Network (RCPN), and from the use of a one-hour basic time-step. However, details of this product may not yet be finalised. For the purposes of assessment of rainfall forecasts, an extra time-delay in the availability of the "merged" product compared to the "quality-controlled" product would be unimportant. One of the features of the "merged" product is that it fully integrates raingauge information into the final product. It is therefore clear that any problems with these raingauge data that cannot be identified using real-time quality-control procedures will be carried over into the final "merged" product and may well have an important effect. Since it is not yet operational, no experience with the properties of this product has been built-up: however, we would expect the same considerations as discussed under (d) to apply. Thus, given that the product provides a useful and stable product for real-time purposes, which implicitly requires that quality-control of the RCPN data has been implemented successfully, locally archived data from the "merged" Nimrod data may be useful for routine monitoring of the rainfall forecasts. The more stringent requirements needed for comparing different variants of forecast procedures may demand further quality control of the data and this would allow the opportunity to bring in data from more extensive sets of telemetered raingauges.

There are a number of other problems relating to these data sources that need to be considered. Firstly, data from recording and telemetered raingauges may well be worthless during snowfall periods unless the instruments are of a specially designed and expensive type, since the snowfall may not be recorded, whereas the melting of snow would be recorded. Quality-control of raingauge data would need to take this possibility into account. Data from daily-read raingauges are notionally not affected by snowfall events because the procedures for recording measurements from such gauges contain explicit provisions for cases where the gauge contains snow. Rainfall values from unadjusted weather radar can be badly affected by "bright band" effects which can lead to substantial over-estimation of rainfall: such effects are most common during periods of cold weather. The problems arising from snowfall and freezing conditions may be such that assessments of forecasts cannot be undertaken for periods where these occur.

4.3 Assessing Accuracy of Performance Measures

4.3.1 Introduction

The procedures for calculating measures of forecast performance that are available in the usual literature, and which have been outlined in Section 2, do not incorporate ways of establishing how well the performance measures are determined by a given dataset. This is often because these performance measures are devised for application on large datasets, which either summarise many forecasts over many forecast-origins, or which combine many sub-area forecasts for a relatively small area over a much larger region.

In principle, there are two somewhat different requirements for measures of accuracy of the performance measures. In the first, a single performance measure is treated, and the concept of accuracy relates to how much different the performance measure might have been if a

different sample of equal size (over a different time-horizon) had been used, or if the performance measure could have been evaluated for an arbitrarily-large dataset. In the second, the performance of forecasts from two sources is to be compared by using the difference between the performance measures for the two forecasts: here attention centres on the accuracy with which the difference can be determined by the available dataset.

The following three sub-sections outline three generally applicable procedures available for assessing whether there is enough evidence to conclude that one forecast-source is better than another. It is arguable that this is the main question to be answered in the present context. One of these procedures is not suited for adoption here because the assumptions on which it is based do not seem applicable. The other two procedures do seem to be useful, although the first can only be used for performance measures that have a certain characteristic structure. Both of these basic procedures have the potential for use in providing an assessment of accuracy of individual performance measures, not just for differences in performance measures. For the present phase of the project, only the first procedure has been applied to the question of determining if there is enough evidence for concluding that one forecast-source is better than another according to a given performance measure.

The procedures here assume that a measure of forecast performance has been selected and that this measure has been evaluated for a number of candidate forecast sources using a standard set of forecast opportunities and observed outcomes. The first of these procedures is based on using the data to estimate the standard deviation of the difference in performance measures. It is only immediately applicable to measures of forecast performance which can be expressed as the average of contributions arising from each forecast occasion. It has the advantage of being able to readily provide feedback, if there were no clear conclusion, on how many forecast occasions would be needed in order to detect an advantage of a given size of one forecast source over another. The second and third procedures can provide an assessment for more general measures of performance and are based on different ways of using resampling methods.

4.3.2 Common Notation for Accuracy Assessment

The description of methods used here is based on an extension of the notation used in Section 2 to define the basic sets of performance measures. For the purposes here, no distinction is made between rainfall amounts and the logarithmic versions of these (which were distinguished by using y or z in Section 2). The symbol y_i is used for the observed value (whether or not a logarithmic or other transformation is used) for a particular instance i, where $i = 1,2,\ldots,n$ indexes the number of occasions for which a comparison can be made. When there are a number of different forecast-sources to consider, it is convenient to distinguish these additional subscripts. For the description here, the forecast-sources will be labelled 1 and 2, although clearly this could be extended to consider a larger number of pairs of forecast sources. Then, corresponding to the observations, two sets of forecast-values are available, denoted by $\hat{y}_{i,1}$ and $\hat{y}_{i,2}$ for $i = 1,2,\ldots,n$.

The measures of forecast performance that are calculated can be considered to be mathematical functions of the observations and forecasts. The values of performance measures will be denoted by T_1 and T_2, where

$$T_1 = G(\{y_i, \hat{y}_{i,1}\}), \qquad T_2 = G(\{y_i, \hat{y}_{i,2}\}),$$

and where $G(\{y_i, \hat{y}_i\})$ denotes the function of the n pairs of observations and forecasts $\{y_i, \hat{y}_i\}$ which defines the performance measure.

4.3.3 Procedure based on estimating the standard error of the mean

The first procedure to be described is based on the usual statistical procedure for estimating the standard error of a mean value. Clearly this procedure can only be applied to those performance measures that are either directly expressed as a mean value, or closely related to such a mean value. In particular, it is assumed that the function defining the performance measure can be expressed in the following way

$$T = G(\{y_i, \hat{y}_i\}) = p\{H(\{y_i, \hat{y}_i\})\},$$

where $p(.)$ is a simple function, and where H is of the special form

$$H = n^{-1}\sum_{i=1}^{n} h(y_i, \hat{y}_i) = n^{-1}\sum_{i=1}^{n} h_i = \bar{h},$$

in which $h_i = h(y_i, \hat{y}_i)$ is specifically a function of the observation and forecast for time-point i only.

For example, the root mean square error is expressible in this form with

$$p(x) = \sqrt{x},$$
$$h(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2.$$

For the comparison of the performance of two forecast sources, the values T_1 and T_2 can be calculated. Essentially the same information is contained in the values obtained before the transformation via the function $p(.)$. Hence the question of whether or not there is enough evidence in the data to say whether one source is better than another can be based on

$$H_1 = n^{-1}\sum_{i=1}^{n} h(y_i, \hat{y}_{i,1}) = n^{-1}\sum_{i=1}^{n} h_{i,1},$$

and

$$H_2 = n^{-1}\sum_{i=1}^{n} h(y_i, \hat{y}_{i,2}) = n^{-1}\sum_{i=1}^{n} h_{i,2}.$$

The difference between H_1 and H_2 can be written as

$$\begin{aligned} H_1 - H_2 &= n^{-1}\sum_{i=1}^{n} h_{i,1} - n^{-1}\sum_{i=1}^{n} h_{i,2}, \\ &= n^{-1}\sum_{i=1}^{n} \{h_{i,1} - h_{i,2}\}, \\ &= n^{-1}\sum_{i=1}^{n} d_i = \bar{d}, \end{aligned}$$

where

$$d_i = h_{i,1} - h_{i,2}, \quad (i = 1, \ldots n).$$

When written in the above form, it can be seen that the difference between the performance measures is simply the mean value of the difference between the per-occasion performance measures expressed by the function $h(.,.)$. Thus the statistic $\bar{d}$ can be referred to as the ***mean difference***. In the present context, it would be expected that there will be statistical dependence between the performance measures $h_{i,1}$ and $h_{i,2}$ relating to the same forecast-occasion, not least because of the functional occurrence of y_i in both

$$h_{i,1} = h(y_i, \hat{y}_{i,1})$$

and

$$h_{i,2} = h(y_i, \hat{y}_{i,2}).$$

However, the difference in performance measures can be expressed in terms of the differences $\{d_i\}$, and it may be reasonable to assume that these terms are statistically uncorrelated across the forecast-occasions. This may be a reasonable assumption provided that the lead-time periods for the forecasts do not overlap, and provided that the target-times for the forecasts are not too close: the spacing necessary here might be shorter than the decorrelation-time of local weather systems because the quantities concerned are essentially differences in forecast errors, not the rainfall quantities themselves. The assumption that the differences $\{d_i\}$ are uncorrelated allows some standard statistical results to be used, but it is one that should ideally be checked. The standard statistical theory indicates that the variance with which $\bar{d}$ estimates the long-term mean difference can itself be estimated by

$$w^2 = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \{ d_i - \bar{d} \}^2.$$

Thus w is the ***standard error of the mean difference***. The statistic that is most directly useful in determining whether there is enough evidence in the data to determine whether one forecast source is better than another is the ***standardised difference***, t, given by

$$t = \bar{d} / w.$$

The way in which the standardised difference is defined can be recognised as being essentially similar to a (paired) Student's *t*-test, and the statistic would have a Student's *t* distribution if certain further assumptions were thought appropriate. In the present context, the assumption that the differences $\{d_i\}$ are uncorrelated, equal-variance and Normally distributed seems unlikely to be tenable. Because of this, the usual hypothesis testing approach is not used (this would use a Student's *t* distribution under the null hypothesis of "no difference in forecast performance"). Instead a somewhat less formal stance is taken and the standardised difference is used as the final indicator of the strength of evidence, with a guide as to the meaning of the values found taken in only a loose sense from the tables of the Student's *t* distribution. Values of the standardised difference that are larger in absolute value than about 2 can be taken as reasonable evidence of a real difference in forecast performance. The sign of the standardised difference would indicate which forecast source is preferred, depending on whether large or small values of the basic performance measure are to be preferred.

Note that the criterion-value "2" for the standardised difference should ideally be increased if the sample size (number of forecast occasions) is smaller than about 20. As already said, there is some underlying theory which suggests the use of a Student's *t*-distribution in determining

this criterion. Thus a formal hypothesis test at the 5% level would use a value of 1.96 for very large sample sizes, 2.00 for a sample size of 60, 2.09 for a sample size of 20, and 2.26 for a sample size of 10. However, there are various assumptions made in the underlying theory which are unlikely to hold in the present circumstances, at least for small sample sizes. If account could be taken of this, it would lead to using still larger values for the criterion-value for small samples. For the informal use of the criterion outlined above, we suggest that the criterion values should be 2.00 for sample sizes above 60 should be used, 2.1 for sample sizes near 20, and 2.5 for sample sizes near 10. This makes some allowance for the problems at small sample sizes and the adjustment will mean that larger values of the standardised difference would be required before the criterion suggests that there is strong evidence for an apparent difference in performance. Ideally, sample sizes should be moderately large. We suggest that, where datasets are selected by first identifying "events", no definitive conclusion about differences in forecast performance should be made based on fewer than four events, where there would be several forecast occasions within each event. Overall, a minimum of 30 forecast occasions should be adequate to avoid problems from the unknown effects of small sample-sizes on the criterion-value.

A similar approach to that above can be applied to just a single performance measure to give a range of values for the performance measure that should cover the long-term average value. Thus, with w_h defined by

$$w_h{}^2 = \{n(n-1)\}^{-1} \sum_{i=1}^{n} \{ h_i - \bar{h} \}^2 ,$$

a range for the long-term value of $\bar{h}$ would be $(\bar{h} - 2w_h, \bar{h} + 2w_h)$, while the corresponding range for the performance measure in its usual (transformed) form would be

$$(p\{\bar{h} - 2w_h\}, p\{\bar{h} + 2w_h\}),$$

where $p\{\bar{h}\}$ is the usual performance measure.

4.3.4 Procedure using permutations

One of the standard ways provided by statistical theory for determining whether there is enough evidence to distinguish two sets of quantities which occur naturally in pairs is a form of permutation test. Here a "pair" refers to the two forecasts from different sources which are available for the same target quantity and for the same forecast occasion. We suggest that this type of permutation test is not suitable for use in this context, but an outline of the procedure is given and we give our reasons for suggesting that this procedure should not be used. The next subsection presents a superficially similar procedure which avoids the pitfalls of the one described here.

A permutation test is way of implementing a hypothesis test within the usual theoretical and statistical framework. The "null hypothesis" here is that, while the long-term performance of two forecast-sources is the same, individual forecasts may differ. However the individual forecasts would only differ in a way that is essentially just random noise and not in ways that are statistically related to the observed outcomes. That is, it should not be the case that one forecast source over-forecasts at low rainfall while the second over-forecasts at high rainfall. Our reason for suggesting that this type of permutation test should not be used for this project is that we consider that this assumption about the statistical behaviour under the null

hypothesis is untenable. The assumption is essentially that the forecasts are interchangeable in a statistical sense.

If the null hypothesis were accepted, this would mean that there is not sufficient evidence to distinguish the two forecast sources in terms of forecast-performance according to the performance-measure being used. If the null hypothesis were rejected, this would indicate that there is sufficient evidence to conclude that the source having the better performance measure for the sample would continue to have better performance over new forecast occasions.

The procedure for implementing the permutation test is as follows.

First evaluate the observed difference in performance, $T_{obs} = T = T_1 - T_2$. This value is based on the dataset actually available, which can be considered as consisting of a number of "triples": $(y_i, \hat{y}_{i,1}, \hat{y}_{i,2})$ for $i = 1, \ldots, n$.

Next evaluate a large number (say 1000) of random samples of T corresponding to the null hypothesis. The k'th such random sample is created as follows:

(i) create a new version of the set of triples: $\left\{\left(y_i^{(k)}, \hat{y}_{i,1}^{(k)}, \hat{y}_{i,2}^{(k)}\right); i = 1, \ldots n\right\}$. This set is created as follows: for each i separately

(a) $y_i^{(k)} = y_i$ always ,

(b) with probability ½: $\hat{y}_{i,1}^{(k)} = \hat{y}_{i,1}$ and $\hat{y}_{i,2}^{(k)} = \hat{y}_{i,2}$;

otherwise : $\hat{y}_{i,1}^{(k)} = \hat{y}_{i,2}$ and $\hat{y}_{i,2}^{(k)} = \hat{y}_{i,1}$.

The effect here is to create an alternative version of the dataset in which the forecasts from the two sources have been swapped at random.

(ii) Calculate the k'th value of the difference in performance:

$$T^{(k)} = T_1^{(k)} - T_2^{(k)},$$

where

$$T_1^{(k)} = G\left(\left\{y_i^{(k)}, \hat{y}_{i,1}^{(k)}\right\}\right), \quad T_2^{(k)} = G\left(\left\{y_i^{(k)}, \hat{y}_{i,2}^{(k)}\right\}\right).$$

(iii) Find what proportion of values $\left\{T^{(k)}\right\}$ are as or more extreme than T_{obs}. This gives the significance level of a two-sided hypothesis test that just accepts the null hypothesis that there is no real difference in the performance of the forecast sources. Alternatively, find the 5%, 10% 90% and 95% points of the empirical distribution of the $\left\{T^{(k)}\right\}$. Possibly also prepare a histogram of the $\left\{T^{(i)}\right\}$ for a display on which the observed outcome T_{obs} would be plotted as a special point. If T_{obs} were between the 5% and 95% points of the empirical distribution of the $\left\{T^{(k)}\right\}$, this would be taken as indicating that there is no strong evidence as to which forecast source has a better long- term performance according to the performance measure being used.

The above description of the permutation test has been outlined in a slightly more complicated way than is strictly necessary in order to facilitate a comparison with the approach described in the next subsection.

4.3.5 Procedure using bootstrapping

Bootstrapping procedures are a further way of implementing a statistical hypothesis test and they can also be used to evaluate the accuracy with which a population statistic is evaluated by a sample of data. Here the assumptions required are that the dataset should consist of a number of statistically independent items. When the purpose is to test whether forecasts from different sources have equivalent performances, the identification of the "triples" made in Section 4.2.4 can be retained and they would potentially be the "items" on which the bootstrapping procedure is based. The underlying notion of bootstrap procedures is that alternative sets of possible datasets can be constructed by randomly selecting among the "items" in the original dataset to form new sample datasets of the same size as the original: the variation between the results found for a collection of such datasets gives an indication of the uncertainty inherent in using the equivalent result from the original dataset. It is clear that for this to be valid, it must be possible to regard the "items" in the original dataset as being "random" in a way that corresponds to the random-selection of items in the bootstrapping procedure. The procedure to be outlined here requires that the items in the original dataset, which are the forecasts and observation for a single forecast-occasion, can be regarded as statistically independent between forecast-occasions. There are variants of the procedure that can allow for certain types of statistical dependence.

The procedure for implementing a bootstrap test is as follows.

It is not necessary to evaluate the observed difference in performance, $T_{obs} = T = T_1 - T_2$. However, the way in which it is constructed is regarded as providing a prototype for evaluating the statistic from other, alternative datasets. Thus the observed difference in performance is considered as a function of the set of "triples": $(y_i, \hat{y}_{i,1}, \hat{y}_{i,2})$ for $i = 1, \ldots, n$.

Next evaluate a large number (say 1000) of random samples of T which are assumed to be statistically equivalent to the original sample. The k'th such random sample is created as follows:

(i) create a new version of the set of triples: $\left\{\left(y_i^{(k)}, \hat{y}_{i,1}^{(k)}, \hat{y}_{i,2}^{(k)}\right); i = 1, \ldots n\right\}$. This set is created as follows: for each i separately,

(a) choose at random and, with equal probability, $j = j_{i,k}$ from among the numbers $1,2,\ldots n$, then

(b) set

$$y_i^{(k)} = y_j, \qquad \hat{y}_{i,1}^{(k)} = \hat{y}_{j,1} \quad \text{and} \quad \hat{y}_{i,2}^{(k)} = \hat{y}_{j,2}.$$

The effect here is to create an alternative version of the dataset in which complete triples have been selected at random, with replacement.

(ii) Calculate the k'th value of the difference in performance:

$$T^{(k)} = T_1^{(k)} - T_2^{(k)},$$

where

$$T_1^{(k)} = G\left(\left\{y_i^{(k)}, \hat{y}_{i,1}^{(k)}\right\}\right), \quad T_2^{(k)} = G\left(\left\{y_i^{(k)}, \hat{y}_{i,2}^{(k)}\right\}\right).$$

(iii) Find whether the set of values $\{T^{(k)}\}$ tend to cover the value zero. The interpretation here is that the values in the set $\{T^{(k)}\}$, taken together, represent the sampling variation inherent in the original sample estimate T_{obs}. Thus, if the values are all on the same side of zero, and well away from zero, this is evidence that the effects of sampling variation cannot be large enough to have caused a situation where the sample-value for the difference has turned out to have the opposite sign from the "true" difference. Thus the apparent preference between the forecast sources shown by T_{obs} would be taken as confirmed. For situations where the conclusion is not so clear-cut, the empirical distribution found from $\{T^{(k)}\}$ provides a direct indication of the size and variation of the difference in performance expected to arise in future when evaluated from new sample datasets of the same size as the original. One way of specifying a confidence interval for the long-term difference in performance would is to find the 5%, and 95% (or 2½% and 97½%) points of the empirical distribution of the $\{T^{(k)}\}$ and to take these as the limits of a 90% (or 95%) confidence interval. If zero were between the 5% and 95% points of the empirical distribution of the $\{T^{(k)}\}$, this would be taken as indicating that there is no strong evidence as to which forecast source has a better long-term performance according to the performance measure being used. The method outlined here corresponds to the bootstrap-percentile method, but there are other methods: the theoretical and practical justifications of the various types of bootstrap-derived confidence intervals are not straightforward and further investigation specific to any particular application would be merited if a bootstrap approach were to be used extensively.

If applied to a single forecast-source, the bootstrap procedure can be used to provide an estimate of the sampling variation in a given sample-statistic. The advantage of this approach over the procedure described in Section 4.2.3 is that it can be applied to a wide range of performance measures, not just those which are expressible as a simple function of a mean value.

4.3.6 Other ways of treating the accuracy question

The problem of taking account of the uncertainty in the value of performance measures can be treated in ways which are rather less formal than those described in the above subsections. These other ways are available where the same performance measure is evaluated for several similar cases. For example, in the present application, forecasts are provided separately for several Areas within a Region and a measure of performance can therefore be evaluated separately for the forecasts for each Area. If there is no reason to suppose that the true forecast-performance will differ radically between the Areas, the variation in the values of the performance measures calculated for the Areas provides a guide to how much the values are affected by sampling variation. A similar argument can be applied to cases where forecast performance is evaluated for a number of different lead-times. Here the true performance

would be expected to change smoothly and to become worse as the lead-time increases. Thus the variation in the calculated performance measures when considered as a function of lead-time provides an indication of how well the true performance has been determined. Conclusions derived in these ways clearly need to be treated with caution.

An apparently important special case arises in the routine monitoring of forecast performance, perhaps on a monthly basis. Here the performance measure would be calculated for a sequence of month-long time-periods. It seems likely that there would be a natural seasonal variation in the predictability of rainfall, but this may not be particularly smooth, given the seasonal nature of synoptic structures. Thus it may be necessary to build-up several years'-worth of experience of forecast performance before much use can be made of the variation between monthly values of the performance measures in indicating sampling uncertainty.

4.4 Summary

Section 4 has discussed the issues related to forecast assessment procedures that arise directly from the specific applications being considered. While assessment procedures have been implemented elsewhere for various types of forecasting problems, each instance has its own requirements. Section 4.1 has categorised some of the issues affecting the choice of an assessment procedure and has discussed how these relate to the requirements of the particular application being considered here. Sections 4.2 and 4.3 go on to discuss, in greater detail, two of the main questions which arise for the present application.

Section 4.2 has discussed the question of defining a suitable ground-truth for use in the assessment procedure. The discussion here has highlighted the point that there are two somewhat distinct requirements for forecast assessments and that these should be treated separately when considering ground-truths. Thus rather more effort in constructing the ground-truth can be justified for one-off studies comparing forecast procedures, where a decision will have a long-lasting effect, than for routine monitoring of forecast performance. A suitable ground-truth can also be affected by the type of forecast target being considered. Where a spatial-average rainfall is the target, a network containing a sufficient number of raingauges can be adequate for routine monitoring of forecasts, although the inclusion of information derived from radar would be regarded as beneficial. Use of a merged radar-raingauge product is suggested as being necessary for comparing different sources of forecasts of spatial average rainfall, particularly where the sources may differ in their treatment of orographic effects or in their use of local knowledge. Where the target of a forecast is the maximum rainfall within a region, use of a merged radar-raingauge product is suggested because typical raingauge networks do not provide sufficient spatial resolution. Some concerns have been raised about the possible problems that might arise from using an archived version of the operational merged radar-raingauge product and, in particular, there are questions of whether there might be a need to re-process the data following subsequent quality control of the raingauge and radar data.

Section 4.3 has discussed some questions arising from the use of relatively small amounts of data in forecast assessment procedures for the present types of application. Many other implementations of assessment procedures have involved circumstances that are considerably more data-rich than the ones of concern here. The main problem arising from the use of small datasets is that any measure of forecast performance will not be very well determined: in other

words it will be subject to random error arising from the sampling of the dataset provided for the assessment. Section 4.3 has discussed some ways in which the sampling error of the performance measures can be estimated. It has gone on to consider the question of determining whether, in the case of comparing two sources of forecasts for the same events, there is enough evidence to conclude that one source is better than the other. This latter question is treated in the results for case studies presented in Section 5.

5. CASE STUDY ASSESSMENT

5.1 Introduction

Section 5 aims to use case study storms selected from across the regions of the Environment Agency to gain experience of the application of the assessment procedure, and associated performance measures, on real data. This experience, together with the review and development of methods for assessment, will be used to make recommendations for operational assessment. Initially the case studies are outlined, including consideration of the "ground truth" rainfall data available to assess the forecasts. Alternative forecast sources, termed Comparative Forecasts, to be used in the assessment are discussed: these include Nimrod radar rainfall forecasts and Mesoscale Model forecasts. The main sub-sections of Section 5 contain detailed assessments of the Daily Weather Forecast, the Evening Update and lastly the Heavy Rainfall Warnings, each with a summary overview.

5.2 The Case Studies

Each of the three different types of rainfall forecasts provided to the Environment Agency by the Met Office relate to a variety of quantities and time-periods. In addition, the target quantities for the rainfall forecasts differ between Regions of the Agency, as do the targeted time-periods and formats of the forecasts. For example, in the case of Heavy Rainfall Warnings, the fact that a warning has been issued may, for some Regions, be the only quantitative information about the expected rainfall amount, while, for other Regions, the warning notice will usually contain an explicit forecast of rainfall amount. For the present study, with its restricted resources, it has been necessary to deal with only a limited number of types and formats of forecasts.

In the preparation for the initial phase of this project, each Region of the Agency was invited to nominate rainfall events that had been notable for their region. Given some overlaps between regions, this gave a set of 11 rainfall events within the calendar year 2002 (up to November), each lasting from 1 to 3 days. This set of events is given in Table 5.2.1. In order to restrict the amount of information being requested, but still to include some instances of forecasts covering periods when there was little or no rainfall, a decision was made to request records of all forecasts made from 5 days before the beginning to one day after each of the initially-identified event periods. The choice of 5 days was related to the lead-time covered by the Daily Weather Forecasts received by most regions, and was defined so that the first Forecast requested would cover a period ending just before the start of the (main) rainfall event. Similarly the last Forecast requested would start on the day after the rainfall event had finished. Because some of the initial set of events were reasonably close together, the time-periods for which forecasts were requested merged into 5 longer case-study periods. These are listed in Table 5.2.2.

Copies of forecasts where the issue date was within the periods indicated in Table 5.2.2 were received for all regions, with the following exceptions.

> **All types of forecasts:** Forecasts issued by London Weather Centre for Anglian, Southern and Thames Regions on 4 February 2002 had not been

Table 5.2.1 Initial set of rainfall events

Event Number	Rainfall Period	EA Description	Region
1	26 January 2002	Frontal	Southern
2	31 January-2 February 2002	Frontal, Cumbria floods	Northwest
3	8-10 February 2002	Frontal	Midlands
4a	14 June 2002	Rapid Thunderstorms	Northeast
4b	16 June 2002	False HRW	Northeast
5a	30-31 July 2002	Convective/Frontal	Midlands
5b	30 July-2 August 2002	N.York Moors, NE Coast	Northeast
6	3 August 2002	Convective, London	Thames
7	9-10 August	NE Coast, Filey & Scarborough floods	Northeast
8	9 September	Frontal/Convective	Thames, Southern, Southwest
9	13 October 2002	Frontal/Widespread	Southwest
10	20-29 October 2002	Over Pennines	Northeast
11	1-3 November 2002	Under-prediction followed by Over-prediction	Northwest

Table 5.2.2 Time-periods for which forecasts were acquired

Period Number	First forecast day	Last forecast day
1	21 January 2002	11 February 2002
2	9 June 2002	17 June 2002
3	25 July 2002	11 August 2002
4	4 September 2002	10 September 2002
5	8 October 2002	4 November 2002

archived by the Met Office. Part of this information (for Thames Region only) has subsequently been received from an Agency source, but too late for inclusion in the present study. It is known that one Heavy Rainfall Warning was issued on this date for Thames Region.

Evening Updates: This type of forecast service was only provided for Thames Region from the beginning of 2002, and for Anglian and Southern Regions from July 2002 onwards.

The forecasts received included revisions made to Daily Weather Forecasts and amendments and cancellations of Heavy Rainfall Warnings.

5.3 Assessment of Daily Weather Forecasts

5.3.1 Approach to Assessment

The Daily Weather Forecasts issued to the eight Environment Agency Regions are of a number of different formats and contain a variety of forecast information. The quantitative rainfall forecast component of the forecasts commonly consists of rainfall quantities forecast for sub-areas of the region for periods of 6 to 24 hours out to a maximum of five days. The specific target forecast quantities, numbers and definition of sub-areas and forecast periods vary across the regions, with some regions receiving forecasts of a single quantity for numerous sub-areas (e.g. Northeast), whilst others receive a forecast of more than one quantity for three sub-areas corresponding to the Agency areas within the region (e.g. Thames, Southern and Anglian).

The availability in an electronic form suited to automated extraction of the forecast quantities had lead to forecasts from Thames, Northwest and Northeast regions being selected for this part of the case study assessment. For Thames Region, assessment using the largest number of comparative forecasts and ground truths was carried out using data supplied for the two events nominated by the region and given in Table 5.2.1, each consisting of 6 forecast occasions. Automated processes were also developed which allowed large numbers of Daily Weather Forecasts to be analysed along with a reduced set of ground truths. A change in the format of the forecasts at the beginning of July 2002 lead to the selection of the forecasts issued in July or later for this part of the assessment. These corresponded to periods 3, 4 and 5 indicated in Table 5.2.2, a total of 53 forecasts. For Northeast and Northwest regions, the Daily Weather Forecasts supplied for single individual case study events (as listed in Table 5.2.1) were assessed. For Northwest region Event 2 was used, giving 8 forecasts occasions. For Northeast Region, Event 5b was used, giving 9 forecast occasions.

Ground Truth

For the case studies presented here, the principal source of "ground truth" data has been derived from the network of telemetering raingauges used for operational flood forecasting within each region. Lists of raingauges located within each forecast area for the three regions were derived using GIS tools. A summary of the network information for each is given in Table 5.3.1.1. Maps of the three regions showing the Daily Weather Forecast areas and gauge networks are given in Figures 5.3.1.1 (a) to (c).

Table 5.3.1.1 Raingauge networks used as source of ground truth for Daily Weather Forecast areas within each region.

Region	Area	Area km^2	Number of Raingauges
Thames	1. Northeast	3224	47
	2. Southeast	3504	28
	3. West	6190	25
Northeast	1. Central and North Pennines	3398	13
	2. Cheviot	2444	7
	3. Moors	1885	6
	4. North East Coast	4145	17
	5. South Pennines	3537	27
	6. Vales and Wolds	6199	13
	7. West Pennines	1815	9
Northwest	1. Cumbria and Pennines North of the Ribble	8198	83
	2. Remainder of Lancashire	2456	32
	3. Greater Manchester, Cheshire and Merseyside	4562	41

The raingauge data were provided as "time of tip" or 15 minute accumulations and were processed to form accumulations over the relevant time period for each gauge, from which spatial averages and maxima were formed. A number of different spatial averages were used across the three regions. These included both conventional averages such as mean and median, in addition to other quantities such as a type of mode and a mean of the non-zero raingauge totals. These two latter forms of spatial average were targeted at specific interpretations of the target forecast quantity in Thames Region. Areal average rainfalls were also calculated using the multiquadric method to interpolate a rainfall surface using all the gauges in each region.

In addition to raingauge data, radar ground truths were derived using the Nimrod quality-controlled 2 km product to form accumulations over the relevant periods, from which spatial averages and maxima were then derived.

Comparative Forecasts

Nimrod and Mesoscale Model data were both used to provide comparative forecasts. Nimrod 5km Forecast Accumulations were available at a 30 minute interval with lead times increasing in steps of 15 minutes out to 6 hours. This dataset was used to provide a comparative forecast for the first period of the Daily Weather Forecasts in

Thames and North East regions. Nimrod forecasts were accumulated over the entire 6 hours of forecast and then processed to form spatial mean and maximum rainfall forecasts. Mesoscale Model forecasts were available at a 6 hourly interval (00,06,12 and 18Z) with lead times increasing in steps of 1 hour out to 48 hours. The 00Z forecast was used to provide comparative forecasts for the first 48 hours of the Daily Weather Forecasts. With a spatial resolution of 11 km it was decided that the model forecasts could not be used to derive spatial maxima, and so the model data was used to derive comparative forecasts for spatial mean rainfall only. In addition to Nimrod and Model forecasts, several types of naive forecasts were also used as comparative forecasts.

Figure 5.3.1.1 a) Thames Region Daily Weather Forecast areas and raingauge network.

Figure 5.3.1.1 (b) Northeast Region Daily Weather Forecast areas and raingauge network.

Figure 5.3.1.1 (c) Northwest Region Daily Weather Forecast areas and raingauge network.

5.3.2 Case Study Assessment for Thames Region

5.3.2.1 Daily Weather Forecast Quantities

The Daily Weather Forecasts issued in Thames Region follow the standard format discussed in Section 3.2.2. An example of a Thames forecast showing the relevant quantitative forecast content is given in Figure 5.3.2.1 Two separate sections of the forecast contain quantitative rainfall amounts. The first for days 1 to 2, gives forecasts of "Amt" and "Max" for each area rainfall for six forecast periods of 6 or 12 hours. Further description of the meaning of these quantities is given below the forecast table. The second section of interest gives forecasts of "Typical Rainfall" and "Most Likely Maximum Rainfall" for 24 hours on days 3, 4 and 5. For this study it has been assumed that "Amt" and "Typical Rainfall" refer to the same target quantity, and the same assumption has been made for "Max" and "Most Likely Maximum Rainfall".

As discussed in Section 3, the precise meaning of the two forecast quantities is somewhat unclear. For this assessment, the quantity referred to as "Max" and "Most Likely Maximum" is assumed to be a spatial maximum of the accumulated rainfall field. The quantity referred to as "Typical Rainfall" could have a number of possible interpretations. As part of this case study assessment, an effort has been made to compare possibilities by considering several alternative forms of ground truth. This has included the specific design of a non-standard "Mode" raingauge quantity, in an attempt to reconstruct one particular interpretation of "Typical Rainfall". This quantity is derived by accumulating the rainfall for each gauge over the appropriate period and then rounding the value for each gauge to the nearest whole number. The mode of the resulting values is then found, with the quantity being treated as missing when there is no single mode. The results of the comparison of different ground truths are given in Section 5.3.2.3. Table 5.3.2.1 summarises the forms of ground truths and comparative forecasts considered for each quantity in the Daily Weather Forecast.

Table 5.3.2.1 Summary of target quantities, ground truths and comparative forecasts for Thames Region Daily Weather Forecasts. Averages refer to spatial averaging carried out on raingauge and radar values which have first been accumulated over the appropriate period.

Quantity: **Typical -**"Amt" and "Typical rainfall" (mm)

Ground truths	Comparative forecasts
Raingauge • Mean • Median • Mode of rounded values • 10 % trimmed Mean • 20 % trimmed Mean • Multiquadric interpolated areal average • Mean of non-zero values **Radar** • Areal average • Median pixel value	**Alternative forecast sources** • Mesoscale model areal average. (Days 1 and 2 only) • Nimrod forecast accumulation areal average. (Day 1 Period 1 only) **Naive forecasts** • Persistence based on previous 6 hours mean raingauge accumulation. • Fixed value of 0 mm. • Fixed value of 0.3mm h^{-1} over the forecast period.

Quantity: **Max** - "Max" and "Most likely maximum rainfall" (mm)

Ground truths	Comparative forecasts
Raingauge • Maximum single gauge **Radar** • Maximum single pixel	**Alternative forecast sources** • Nimrod forecast accumulation spatial maximum. (Day 1 Period 1 only) **Naive forecasts** • Persistence based on previous 6 hours maximum raingauge accumulation. • Fixed value of 0 mm. • Fixed value of 0.3mm h^{-1} over the forecast period.

Day and Date		Amt mm	Cnf	Max mm	Amt mm	Cnf	Max mm	Amt mm	Cnf	Max mm
Tue 08/10/02	0600-1200	0	H	0.5	0.5	H	1	0.5	H	2
	1200-1800	0	H	0.5	0.5	H	2	1	M	3
	1800-2400	0	H	0	0	H	0.5	0	H	0.5
Wed 09/10/02	0001-0600	0	H	0	0	H	0	0	H	0
	0600-1200	0	H	0	0	H	0	0	H	0
	1200-2400	0	H	0	0	H	0	0	H	0

Notes:
Amt:- A typical value of measured rainfall over the Area during the period.
Cnf: A measure of the likelihood of this value being achieved anywhere in the Area in this time period. Guidelines H=more than 60% M= 30-60% L= less than 30%
Max: An indication of the most likely maximum rainfall at any one location in this time period. This is not an extreme value.

Forecast for days 3-5

Forecast for 0001 to 2400 Thursday 10/10/02:

Mostly dry with sunny spells. Increasing cloud in the south may bring the risk of a few showers here later (30%). Fresh easterly winds, 18-22mph with gusts 35mph.

Typical rainfall mm	Most likely maximum rainfall mm	Max temp 0C	Land Wind Dir/mph
0	1	15	E / 20

Forecast for 0001 to 2400 Friday 11/10/02:

Risk perhaps of a few showers in the south at first, otherwise dry with some sunshine. Thickening cloud may also bring some rain to western parts by late evening. Winds easing to a gentle east to southeasterly, 8-12mph.

Typical rainfall mm	Most likely maximum rainfall mm	Max temp 0C	Land Wind Dir/mph
0	2	15	E-SE / 12

Forecast for 0001 to 2400 Saturday 12/10/02:

Mostly cloudy with outbreaks of rain edging slowly east across the region, some heavier bursts possible. Gentle south to southeast winds, veering northwest as the rain clears, 8-12mph.

Typical rainfall mm	Most likely maximum rainfall mm	Max temp 0C	Land Wind Dir/mph
2	5	16	S-SE / 10

Figure 5.3.2.1 Sections of Daily Weather Forecast for Thames Region containing quantitative rainfall forecasts:- Top: Days 1 to 2, Bottom: Days 3 to 5.

5.3.2.2 Basic Statistics of Case Study Data

An initial statistical analysis of the datasets to be used in the case study analysis was carried out. As previously stated in Section 5.3.1, automated procedures allowed a subset of ground truths to be analysed along with the Daily Weather Forecasts for 53 occasions corresponding to all the case study days after 1 July 2002. All ground truths with the exception of multiquadric interpolated raingauge, mean non-zero raingauge (which was added later) and median radar were included in this part of the assessment. The remainder of the ground truths and the comparative forecasts were only analysed for the two identified Thames events in August and September 2002.

Figures 5.3.2.2 (a) to (f) illustrate the mean, median and standard deviation of the forecasts and ground truths used for the "Typical" rainfall quantity for Thames Northeast Area. Figures 5.3.2.2 (g) to (i) illustrate the same statistics for the "Max" rainfall quantity.

Figures 5.3.2.2 (a) to (c) indicate that as expected, the rainfall amounts increase with forecast period (which themselves increase from 6 hours to 24 hours as lead time increases). While the standard deviation of the rainfall amounts is broadly similar for the Daily Weather Forecast Typical Rainfall and the ground truth quantities, the mean and median statistics show that on average, the Daily Weather Forecast Typical Rainfall tends to give higher rainfall values that any of the forms of ground truth. This is the case across all three areas of Thames region, although only the plots for Northeast Area are shown here. Comparing the difference ground truth quantities, the radar tends to give the highest rainfall amounts and mode raingauge the lowest, with the other forms of ground truth tending to appear in the same order between these two extremes.

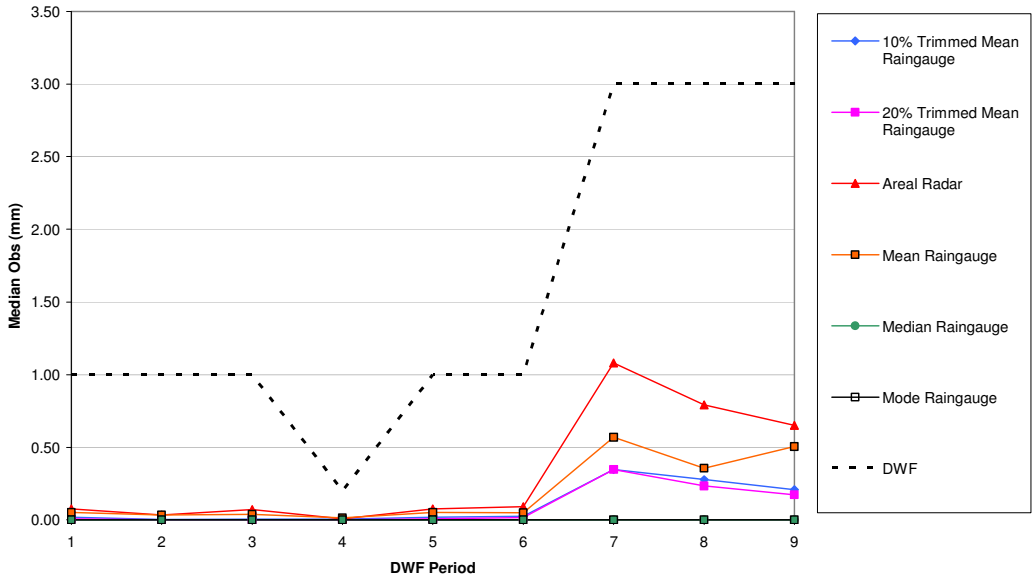
Because of the tendency for the Daily Weather Forecast to overestimate the rainfall amounts according to the results presented in Figures 5.3.2.2 (a) to (c), a further form of ground truth, the mean of non-zero raingauge accumulations, was introduced for the event-only assessment shown in Figures 5.3.2.2 (d) to (f). This form of ground truth replaces the 10% trimmed mean quantity which appears to be similar to the 20% trimmed mean quantity in the assessment as shown in Figures 5.3.2.2 (a) to (c). In Figures 5.3.2.2 (d) to (f) the statistics of all the forecast quantities and ground truths used for the event-only assessment are presented. The figures show that the newly introduced mean of non-zero raingauge ground truth gives higher values which in one case, for Period 2 shown in Figure 5.3.2.2 (d) , is fairly close to the Daily Weather Forecast amount. For most of the other periods shown in Figures 5.3.2.2 (d) and (e), the new ground truth gives values similar to the radar areal average, but not as large as the Daily Weather Forecast "Typical" amount.

Also shown in Figures 5.3.2.2 (d) to (f) are the statistics for the Mesoscale Model forecasts (Periods 1 to 6 only) and Nimrod Forecast Accumulations (Period 1 only). Figures 5.3.2.2 (d) and (e) show the mean and median values of these quantities tend to be closer to those of the ground truths than the Daily Weather Forecast Quantities are, except for one case for the Mesoscale Model for Period 2 shown in 5.3.2.2 (d).

Figures 5.3.2.2 (g) to (i) illustrate the statistics for the "Max" quantity and the two ground truths associated with it. Figure 5.3.2.2 (g) and (i) indicate that the maximum radar accumulations are likely to be affected by anomalous high pixel values, hence the mean and standard deviation for this ground truth are higher than that for raingauges or the Daily Weather Forecast maximum. The median values shown in Figure 5.3.2.2 (h) indicate that, if anomalous high radar accumulated pixels are ignored, the Daily Weather Forecast maximum tends to be closer to the radar than the raingauge ground truth.

A scatter plot of Daily Weather Forecast "Typical" rainfall versus areal radar ground truth is given in Figure 5.3.2.2 (j). The plot shows the data points for all 53 assessment occasions and for all three sub-areas of Thames region. In Figure 5.3.2.2 (k) the log transformed forecast and observed values are plotted, where the revised log transform described in Section 2.2.2 has been used. Comparing Figure 5.3.2.2 (j) and (k), it can be seen that values of exactly 1 mm are transformed to zero, and the smallest non-zero Daily Weather Forecast amounts of 0.2 mm are transformed to approximately -1.6. The effect of the log transformation on large errors can be seen for the point located at (3,17) in Figure (j) which is transformed to approximately (1.1,2.8) in Figure (k). Forecast and observed values less than the threshold of 0.2 mm are transformed to a value of approximately -2.3 and the points corresponding to these values can be clearly seen in Figure (k). Overall the spread of data points on Figure 5.3.2.2 (k) suggests that performance measures making use of the log transformation may put too great an emphasis on errors occurring when forecast or observed values are below the threshold of 0.2 mm.

(a) Mean of ground truth and DWF Typical Rainfall quantities across 53 assessment occasions.



(b) Median of ground truth and DWF Typical Rainfall quantities across 53 assessment occasions.

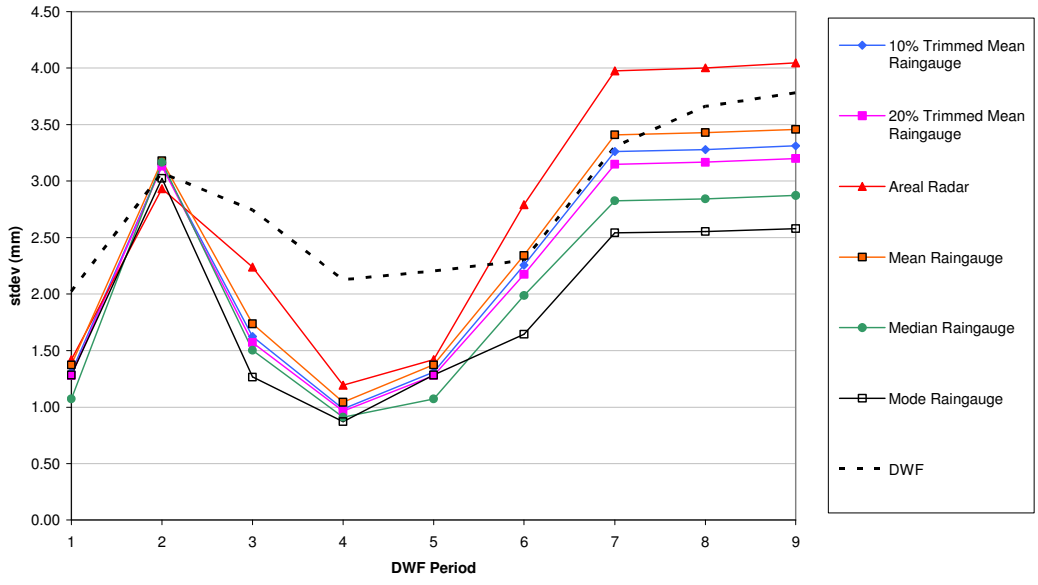
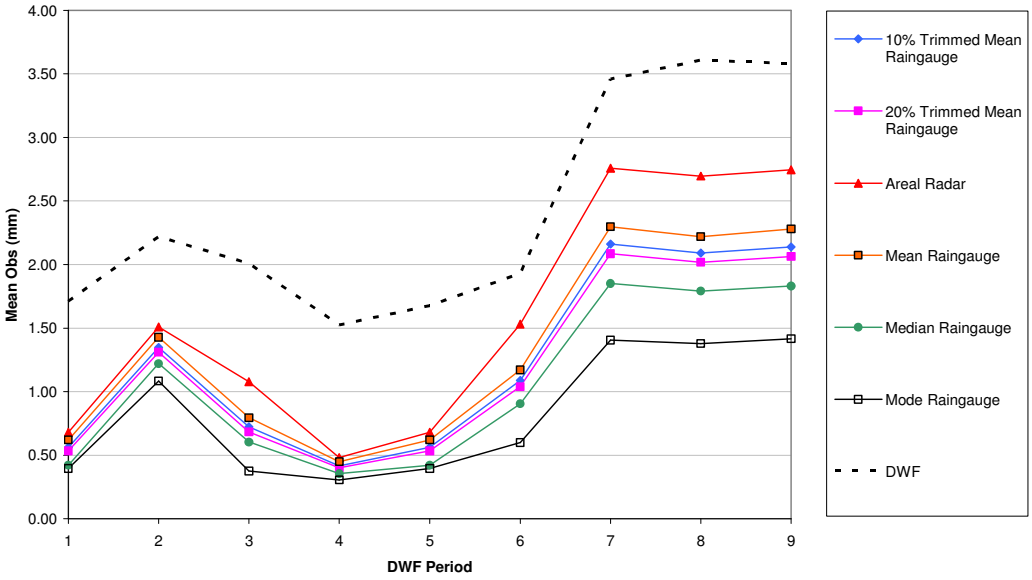


Figure 5.3.2.2 Basic statistics of datasets used for case study assessment Thames Northeast Area.

(c) Standard deviation of ground truth and DWF Typical Rainfall quantities across 53 assessment occasions.



(d) Mean of all ground truths and forecasts considered for Typical Rainfall across two case study events (12 assessment occasions)

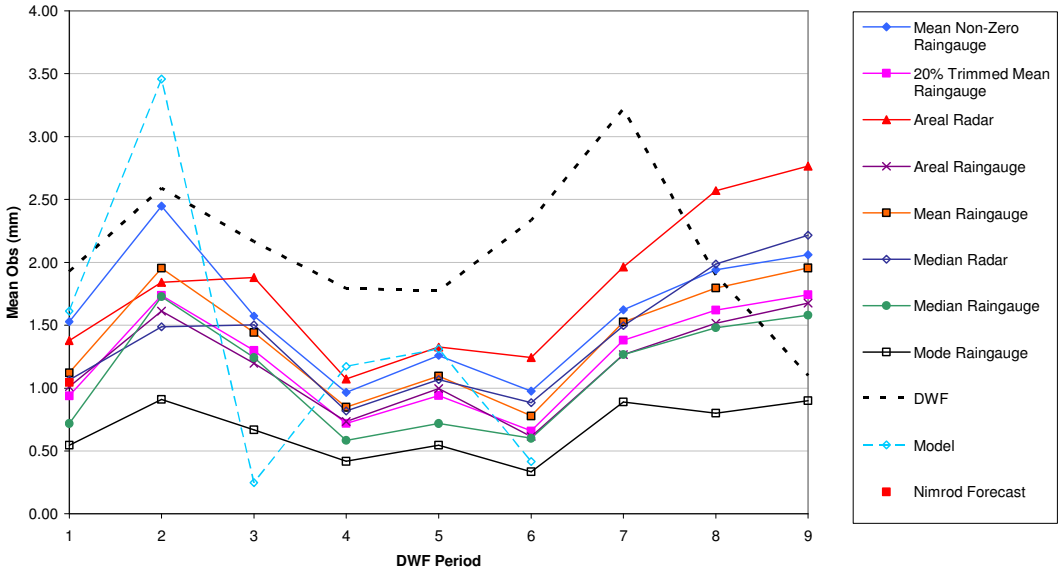
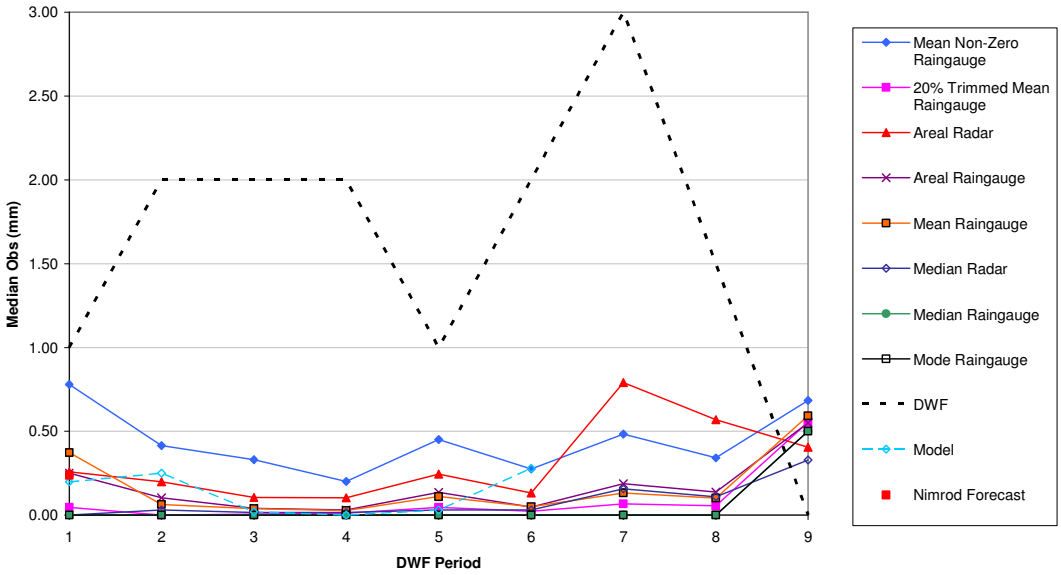


Figure 5.3.2.2 cont' Basic statistics of datasets used for case study assessment Thames Northeast Area.

(e) Median of all ground truths and forecasts considered for Typical Rainfall across two case study events (12 assessment occasions)



(f) Standard Deviation of all ground truths and forecasts considered for Typical Rainfall across two case study events (12 assessment occasions)

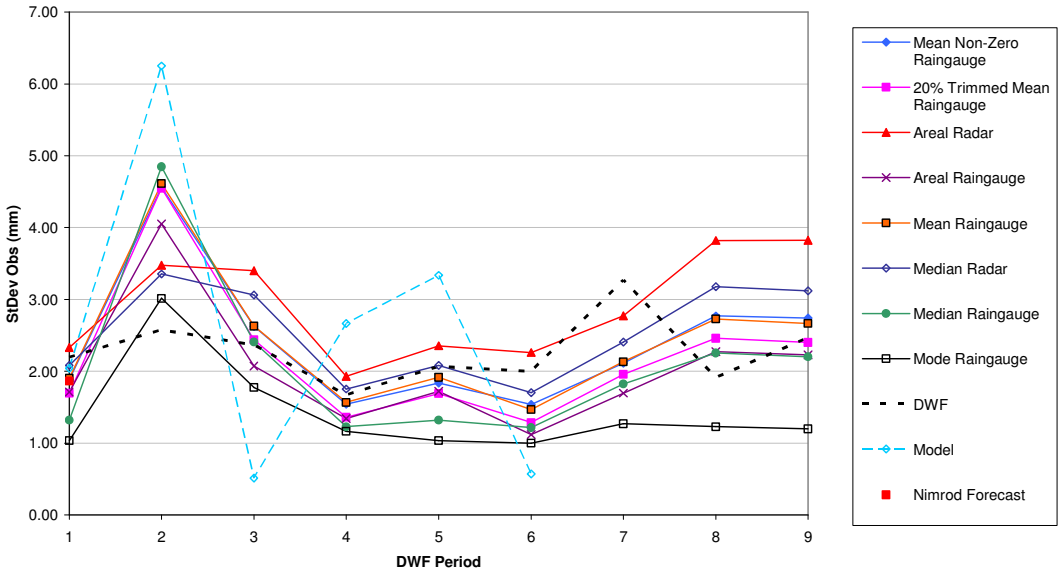
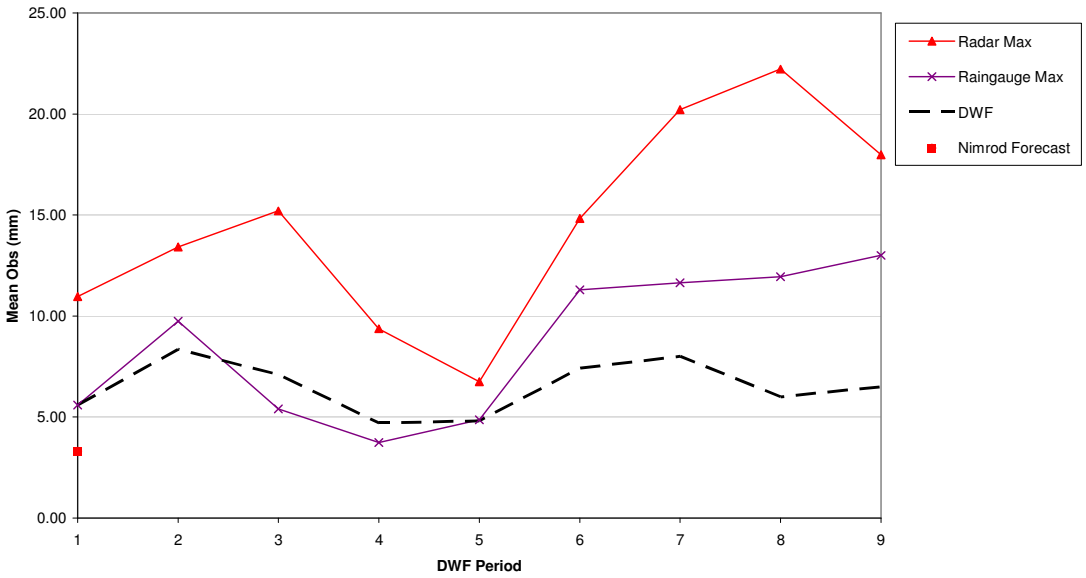


Figure 5.3.2.2 cont' Basic statistics of datasets used for case study assessment Thames Northeast Area.

(g) Mean of all ground truths and forecasts considered for Most Likely Maximum Rainfall across two case study events (12 assessment occasions)



(h) Median of all ground truths and forecasts considered for Most Likely Maximum Rainfall across two case study events (12 assessment occasions)

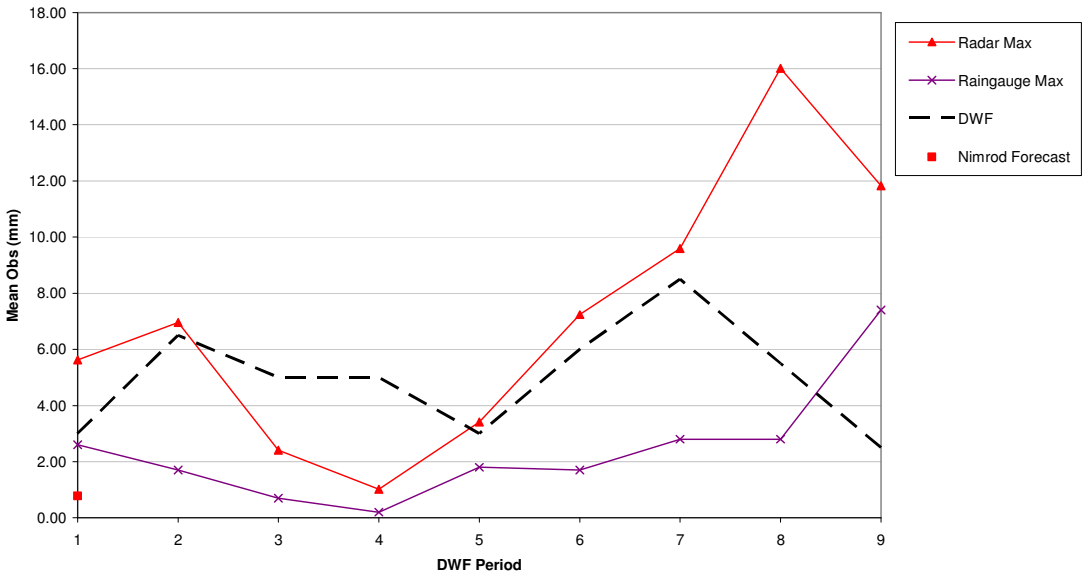
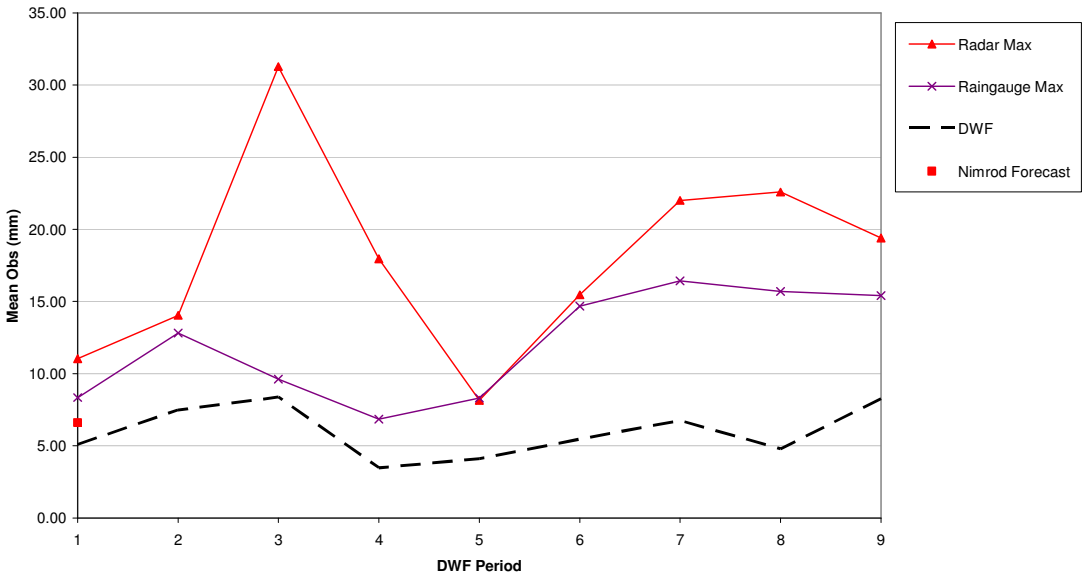


Figure 5.3.2.2 cont' Basic statistics of datasets used for case study assessment Thames Northeast Area.

(i) Standard Deviation of all ground truths and forecasts considered for Most Likely Maximum Rainfall across two case study events (12 assessment occasions)



(j) Scatter plot showing Daily Weather Forecast "Typical" rainfall versus areal radar ground truth for all three Thames sub-areas (53 assessment occasions)

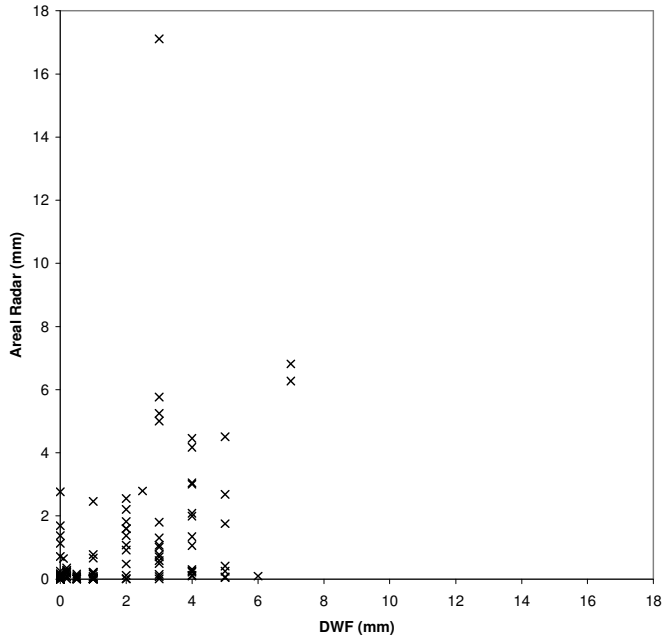


Figure 5.3.2.2 cont' Basic statistics of datasets used for case study assessment

(k) Scatter plot showing revised log-transform of Daily Weather Forecast "Typical" rainfall versus revised log-transform of areal radar ground-truth for all three Thames sub-areas (53 assessment occasions). 39 % of the forecasts and 62 % of the observations fell below the threshold of 0.2 mm

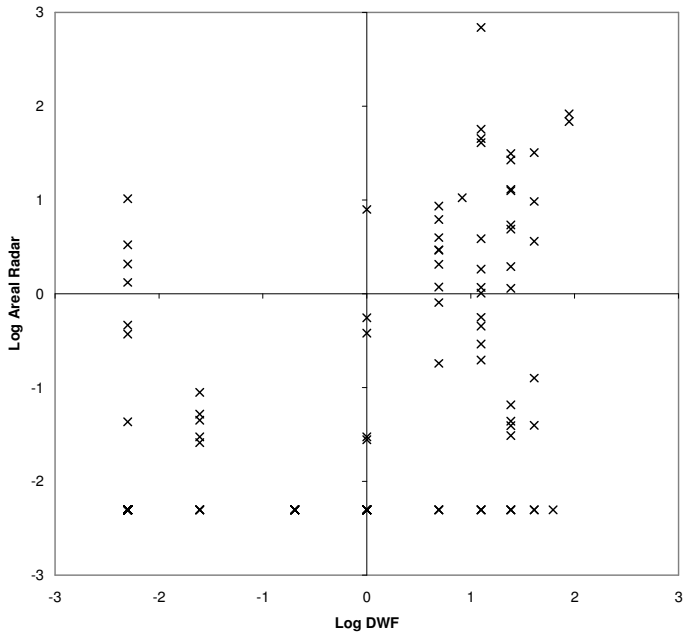


Figure 5.3.2.2 cont' Basic statistics of datasets used for case study assessment.

5.3.2.3 Selection of suitable forms of ground truth

As described in Sections 5.3.2.1 and 5.3.2.2, the uncertainty in the meaning of the "Typical" rainfall quantity lead to a total of ten different possible forms of ground truth quantities derived from raingauges and radar. Before continuing to look at the different performance measures to be applied to the Daily Weather Forecasts, an attempt was made to reduce this set to a manageable number and determine if any of these quantities was more appropriate than the others.

The basic statistics of forecasts and observations presented in Section 5.3.2.2 indicated that on average the Daily Weather Forecast "Typical" rainfall quantity was higher than the forms of ground truth chosen. They also indicated that the different forms of ground truth tend to appear in the same order in terms of the rainfall amount, which suggests the set of ground truths can be reduced to a smaller representative set.

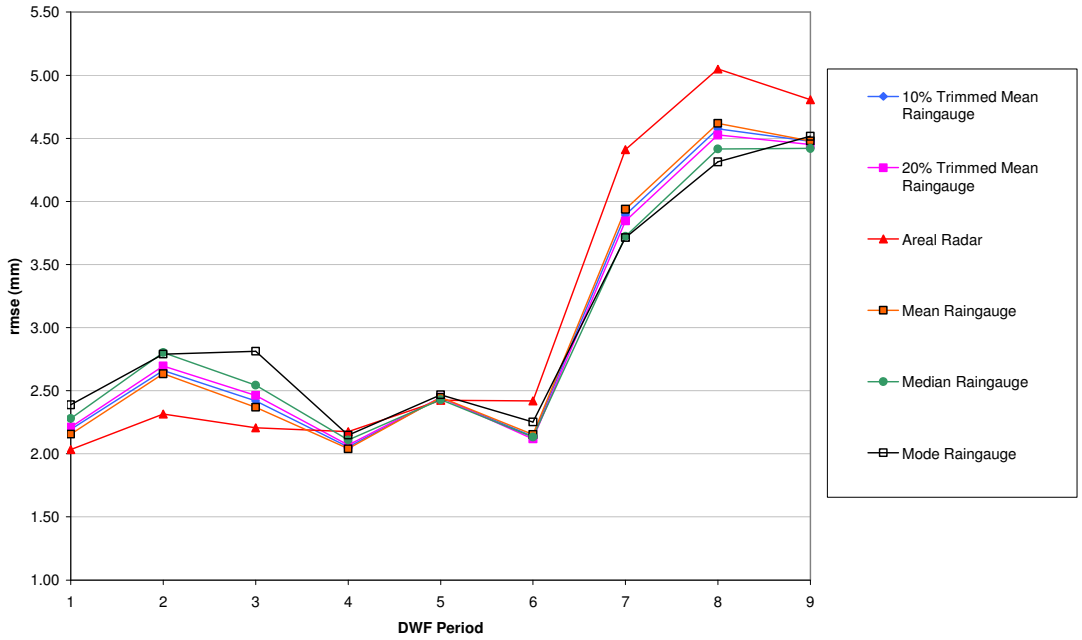
Figures 5.3.2.3 (a) to (f) illustrate four raw performance measures derived for the Daily Weather Forecast "Typical Rain" using the various forms of ground truth. These results suggest that the two extremes in performance can be obtained using the radar areal average and mode raingauge ground truths, with other ground truths usually giving performance measures between these two extremes.

The results presented in Section 5.3.2.2 and Figure 5.3.2.3 lead to the conclusion that no single form of spatial averaging tested here is obviously more closely related to the Daily Weather Forecast "Typical" rainfall quantity than any other. It is recommended that a reduced set of ground truths is chosen for further analysis that best represents the variation encountered here.

Based on these results, the recommended choice for a satisfactory set of ground truths is the following four: Areal Radar, Mean Raingauge, Mean Non-Zero Raingauge and Mode Raingauge. This retains the spread of amounts shown in Section 5.3.2.2 whilst also retaining two independent sources of ground truth (raingauge and radar), and both simple and more complicated methods of deriving the ground truth quantity. However, there may be the need to reduce the set further, in which case just the Areal Radar and Mode Raingauge could be used.

As discussed in Section 5.3.2.2 the basic statistics presented for the "Max" rainfall quantity indicate that the radar ground truth may be prone to anomalous high values. However it can also be argued that in convective events a typical raingauge network may be unable to measure the spatial maximum rainfall accumulation accurately, especially for shorter accumulation periods. It therefore seems sensible to retain both forms of ground truth for the "Max" rainfall quantity.

(a) Root mean square error for ground truths available on 53 assessment occasions.



(b) Mean absolute error for ground truths available on 53 assessment occasions.

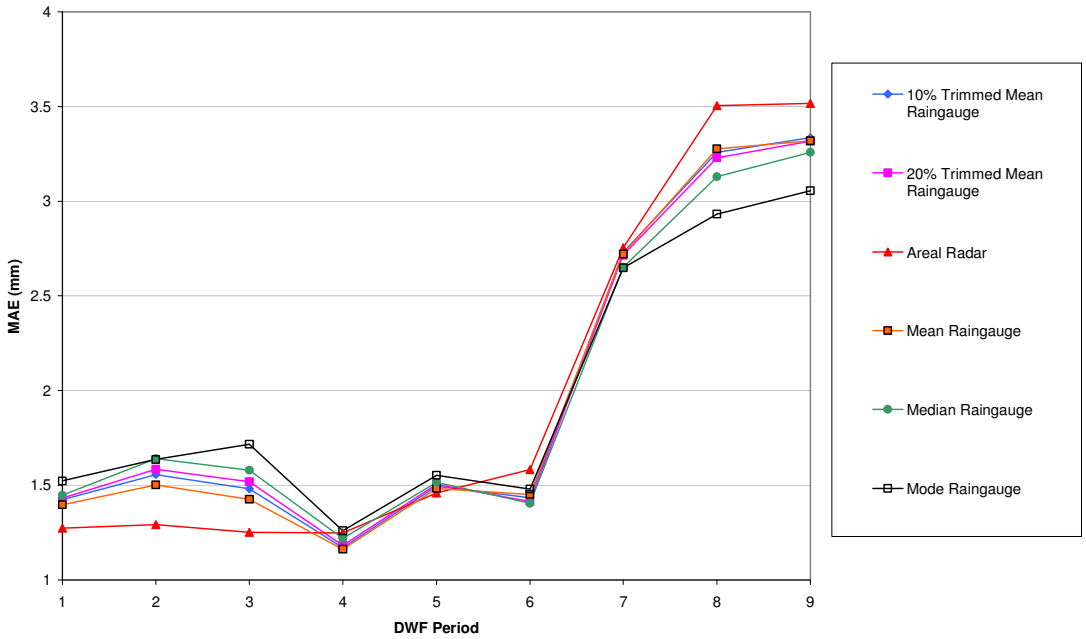
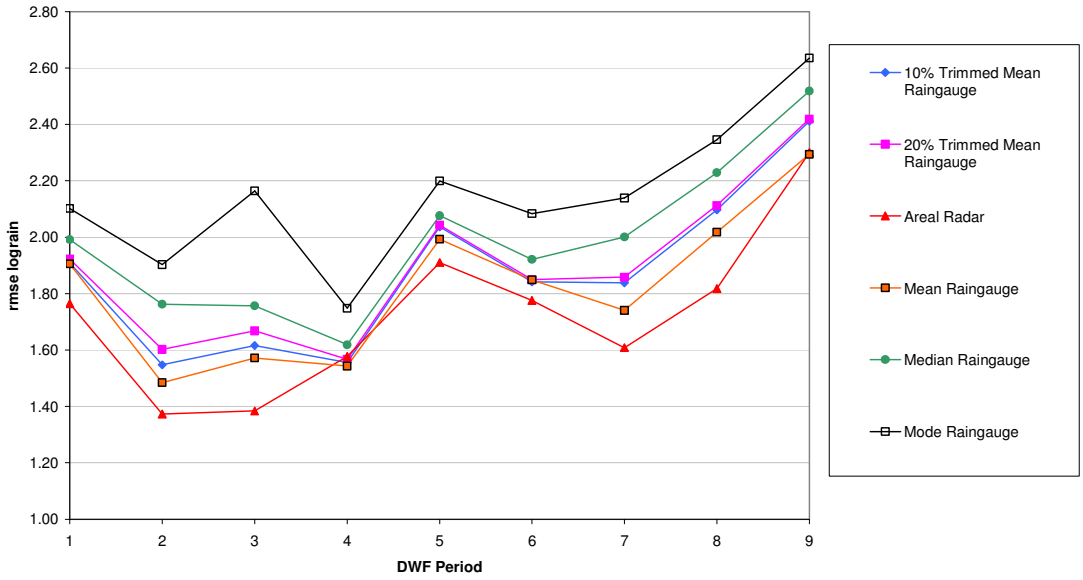


Figure 5.3.2.3 Raw performance measures of Daily Weather Forecast "Typical Rainfall", Thames Northeast Area, obtained using various forms of ground truth.

(c) Root mean square error of log rainfall for ground truths available on 53 assessment occasions.



(d) Mean absolute error of log rainfall for ground truths available on 53 assessment occasions.

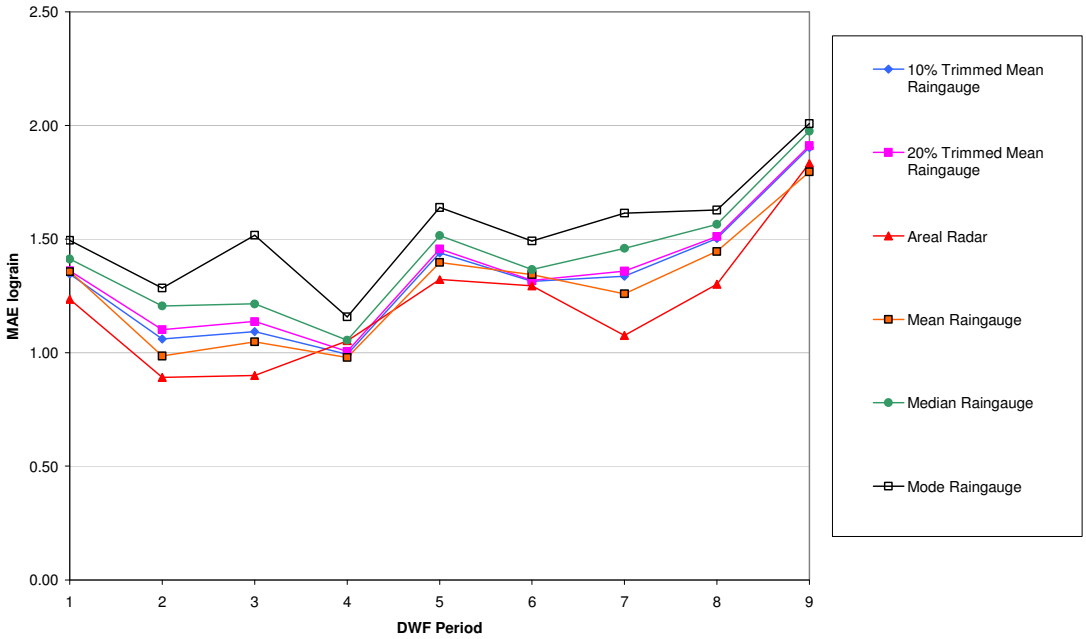
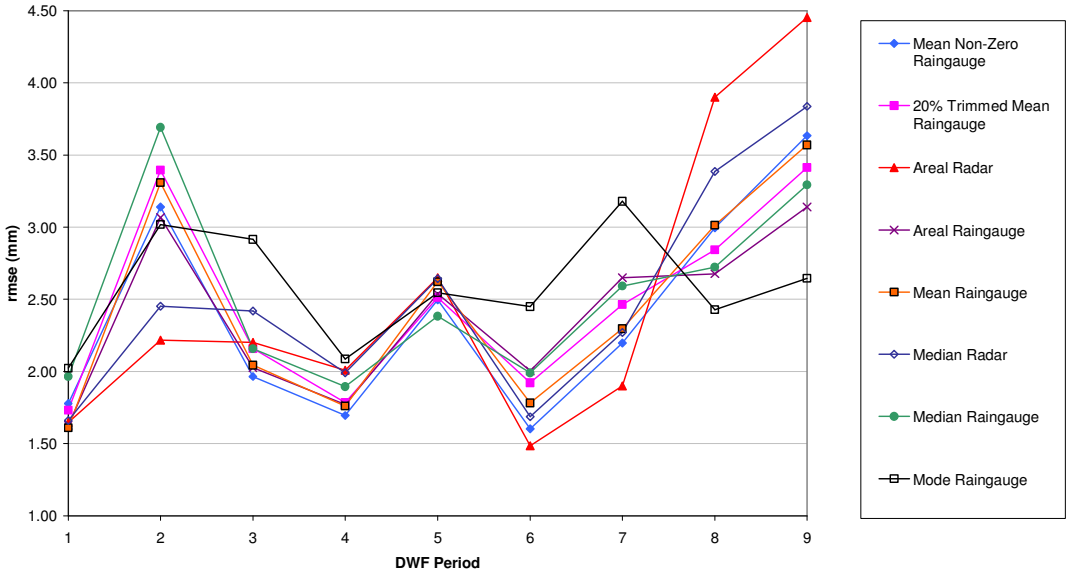


Figure 5.3.2.3 cont' Raw performance measures of Daily Weather Forecast "Typical Rainfall", Thames Northeast Area, obtained using various forms of ground truth.

(e) Root mean square error for ground truths available for two case study events (12 assessment occasions).



(f) Mean absolute error for ground truths available for two case study events (12 assessment occasions).

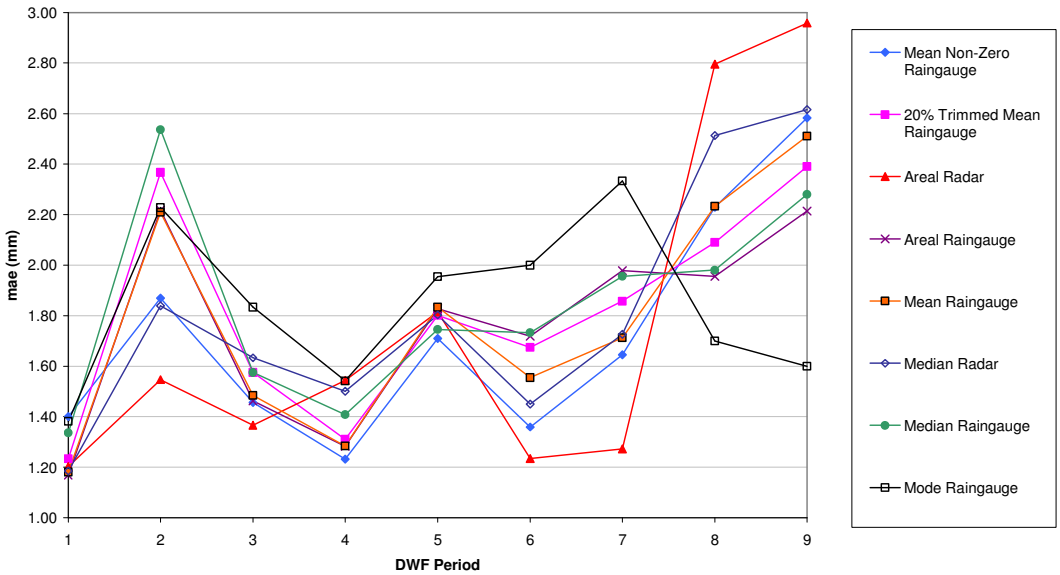


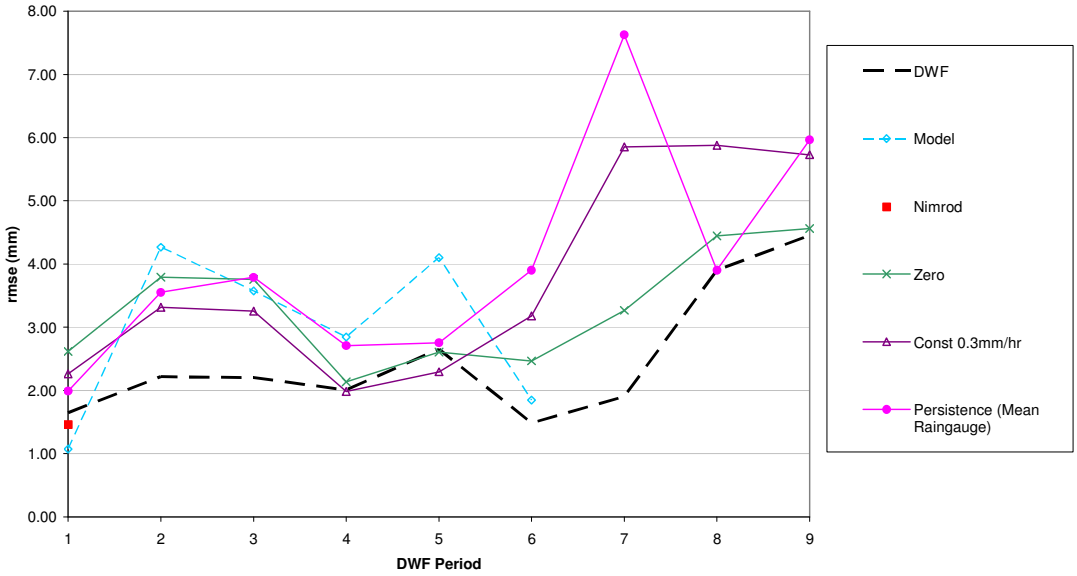
Figure 5.3.2.3 cont' Raw performance measures of Daily Weather Forecast "Typical Rainfall", Thames Northeast Area, obtained using various forms of ground truth.

5.3.2.4 Raw Assessment Measures

Figures 5.3.2.4 (a) to (d) present the root mean square error, mean absolute error , root mean square error of log rainfall and mean absolute error of log rainfall for the "Typical" rainfall quantity, for 12 case study assessment occasions, using the radar areal average ground truth. Figures 5.3.2.4 (e) to (h) present the same statistics obtained using the mode raingauge ground truth.

The most striking feature of these plots is the difference in apparent relative performance of forecasts as computed by the normal and log versions of both root mean square error and mean absolute error performance measures. The normal versions imply that the performance of the Daily Weather Forecasts is similar or better than the Mesoscale Model, whilst the log versions all suggest that the Model performance is better. This implies that there are a few very large errors in the Model forecasts, with other errors being relatively small compared to those of the Daily Weather Forecasts. The log version of the performance measures would reduce the effect of these large errors and hence show the Model to be performing better. Alternatively, the Daily Weather Forecasts may have proportionately large errors during periods of low rainfall compared with those from the Mesoscale Model: The log version of the performance measures would amplify the effect of these errors and hence show the DWFs to be performing worse.

(a) Root mean square error using areal radar ground truth



(b) Mean absolute error using areal radar ground truth

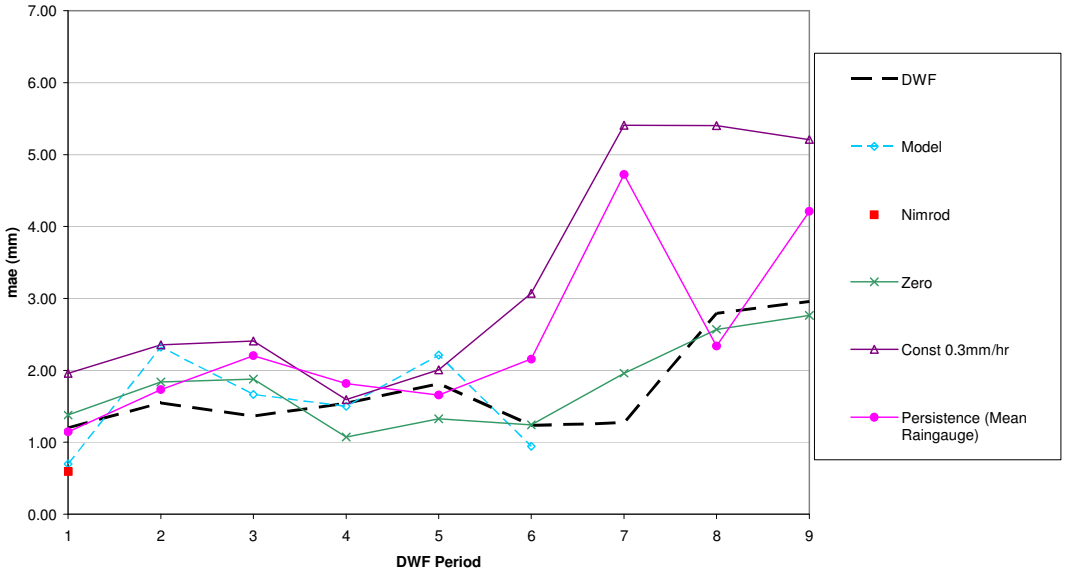
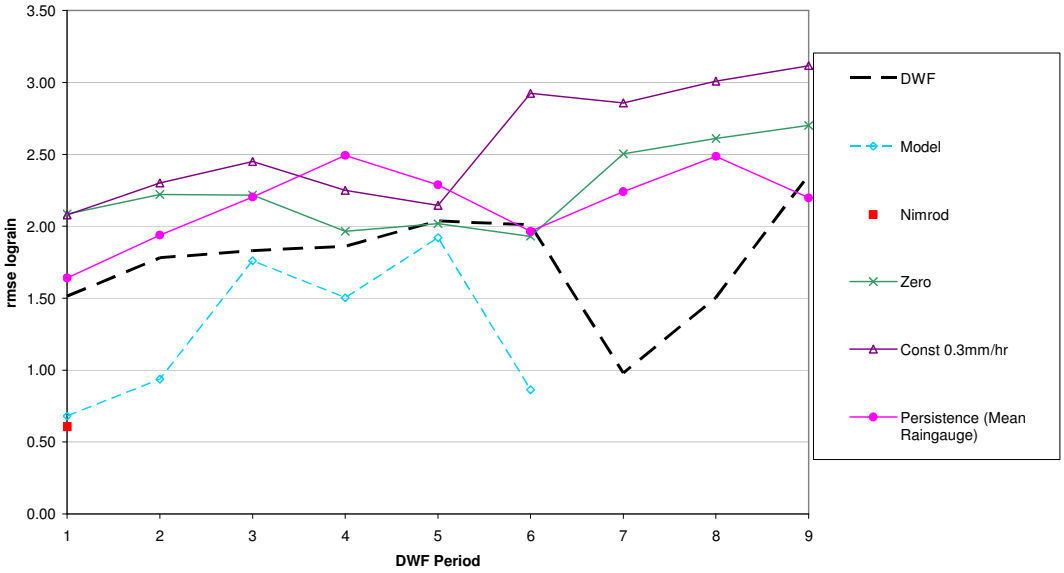


Figure 5.3.2.4 Raw performance measures of Daily Weather Forecast "Typical" rainfall and comparative forecasts, Thames Northeast Area, obtained using radar areal average and modal raingauge ground truths for two case study events (12 forecast occasions).

(c) Root mean square error of log rainfall using areal radar ground truth



(d) Mean absolute error of log rainfall using areal radar ground truth

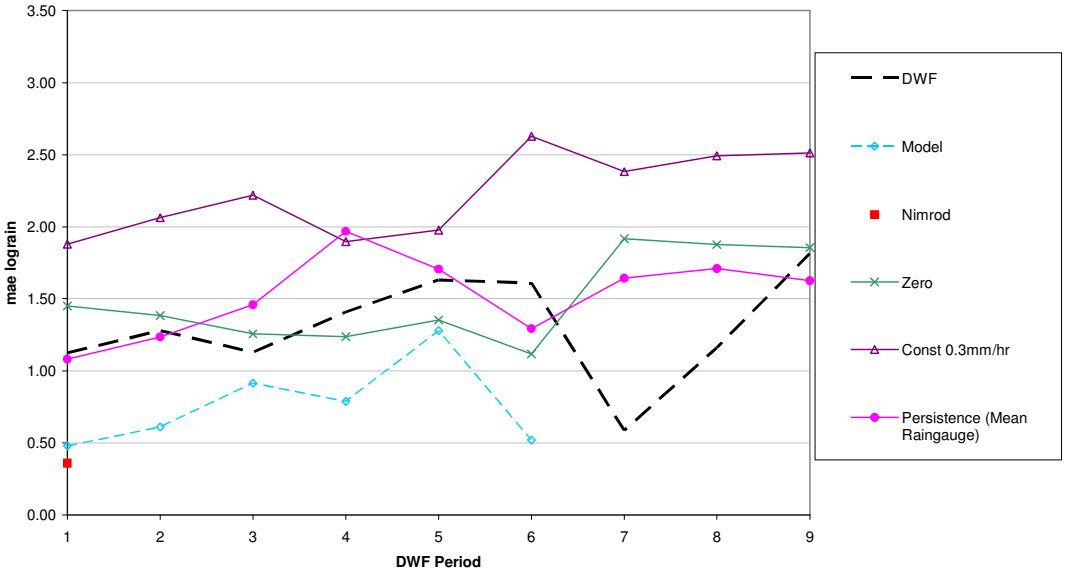
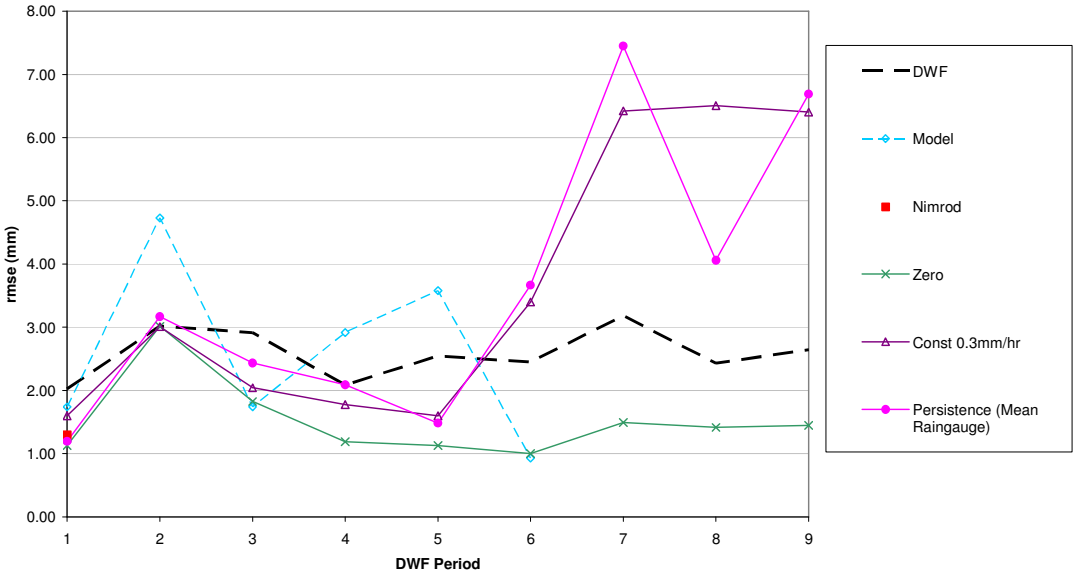


Figure 5.3.2.4 cont' Raw performance measures of Daily Weather Forecast "Typical" rainfall and comparative forecasts, Thames Northeast Area, obtained using radar areal average and modal raingauge ground truths for two case study events (12 forecast occasions).

(e) Root mean square error using modal raingauge ground truth



(f) Mean absolute error using modal raingauge ground truth

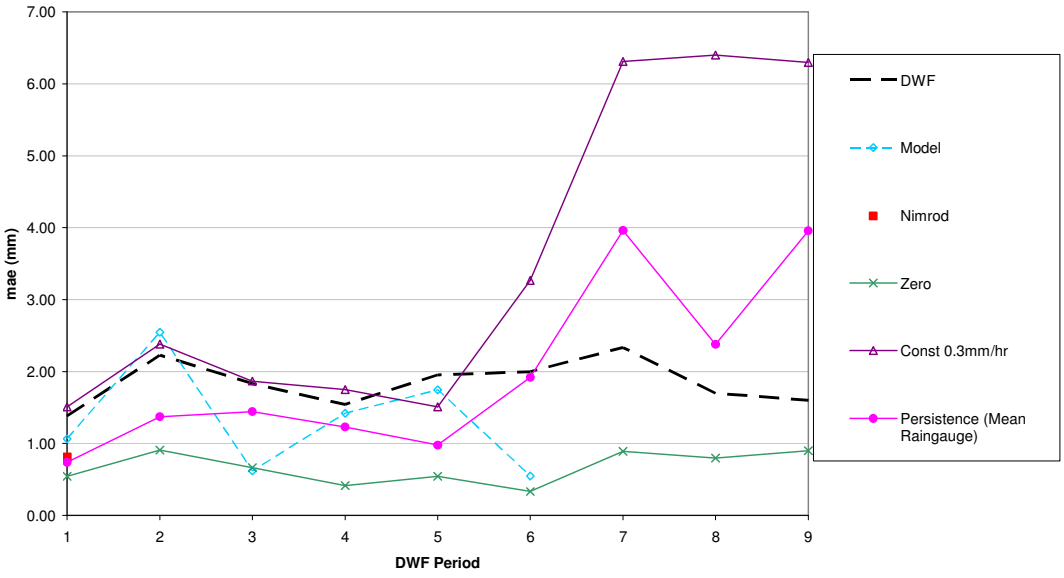
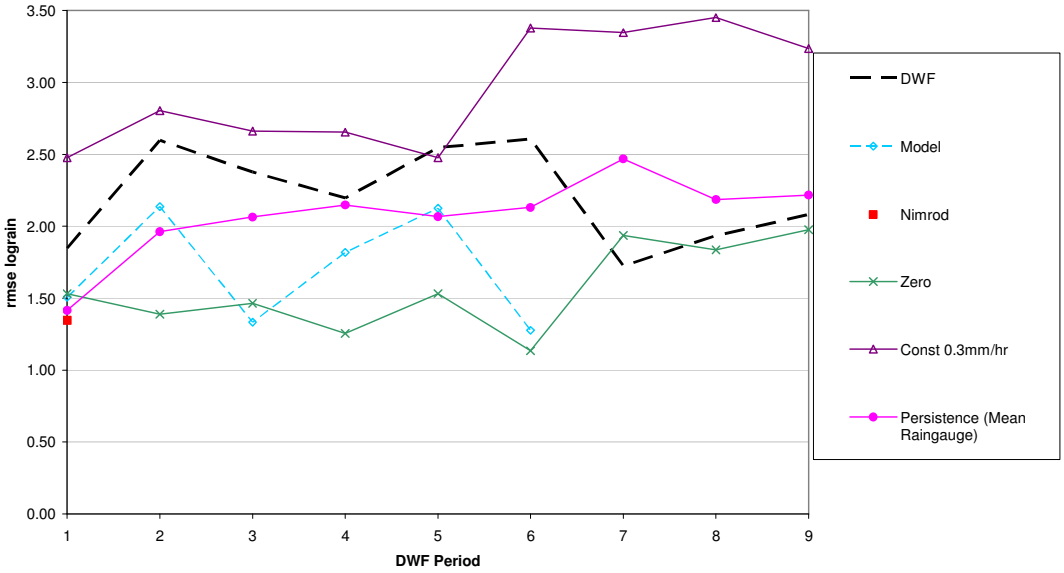


Figure 5.3.2.4 cont' Raw performance measures of Daily Weather Forecast "Typical" rainfall and comparative forecasts, Thames Northeast Area, obtained using radar areal average and modal raingauge ground truths for two case study events (12 forecast occasions).

(g) Root mean square error of log rainfall using modal raingauge ground truth



(h) Mean absolute error of log rainfall using modal raingauge ground truth

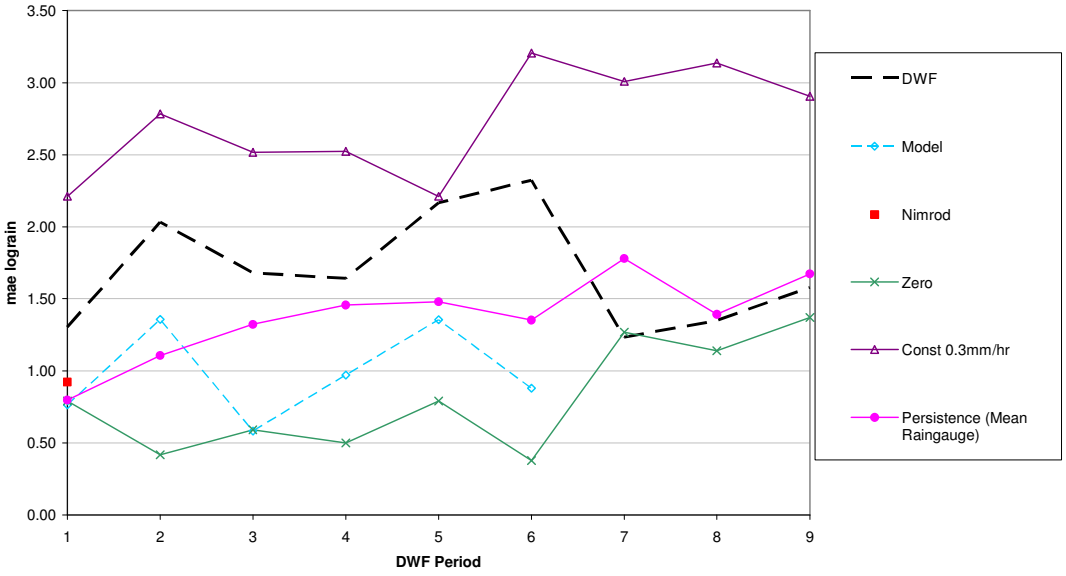


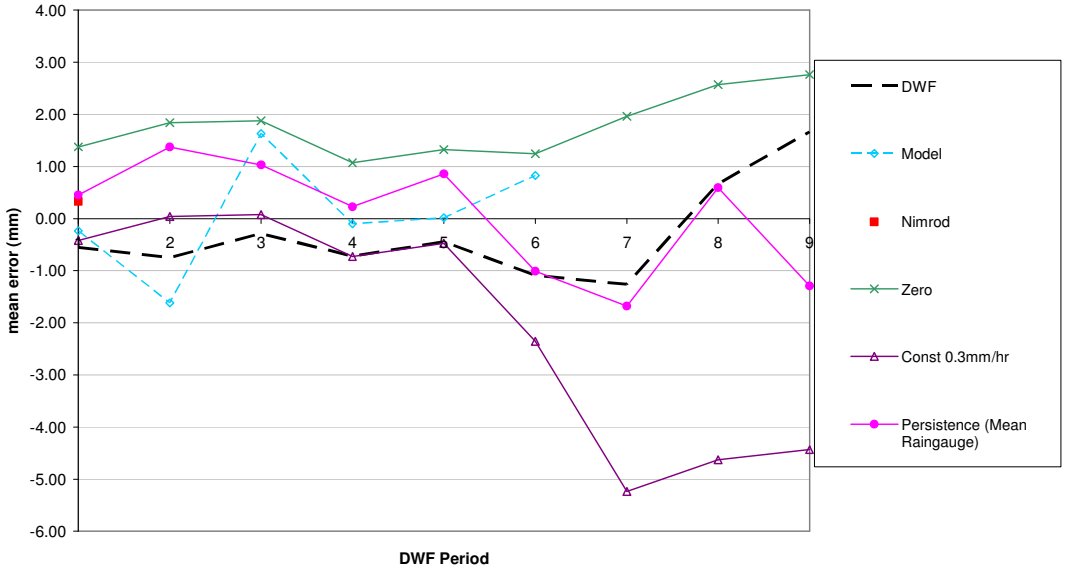
Figure 5.3.2.4 cont' Raw performance measures of Daily Weather Forecast "Typical" rainfall and comparative forecasts, Thames Northeast Area, obtained using radar areal average and modal raingauge ground truths for two case study events (12 forecast occasions).

5.3.2.5 Measures of Bias

Figure 5.3.2.5 shows results for two bias measures - the mean error of rainfall and mean error of log rainfall- for forecasts of the "Typical" rainfall quantity, using radar areal average and mode raingauge forms of ground truth. In these figures a negative error indicates an overestimation of rainfall.

The figure, which shows calculated measures of bias for all 12 forecast occasions, further illustrates the overestimation of rainfall by the Daily Weather Forecasts as discussed in previous sections. The mean error of log rainfall presented here uses the threshold method to deal with small rainfall quantities. It takes into account the magnitude of the rainfall amount, and reduces the relative effect of large errors.

(a) Mean error using radar areal average ground truth



(b) Mean error of log rainfall using radar areal average ground truth

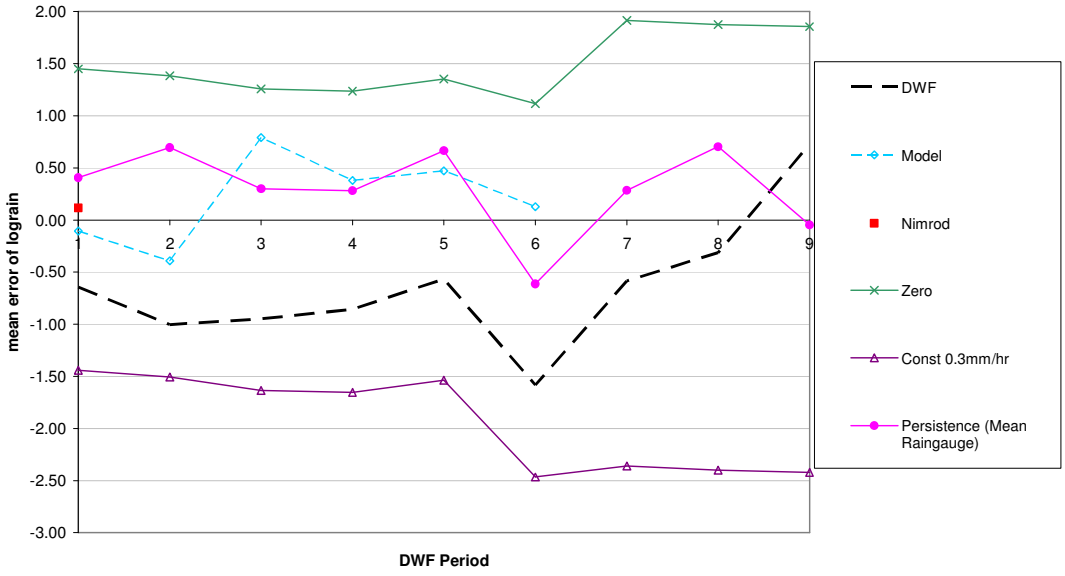
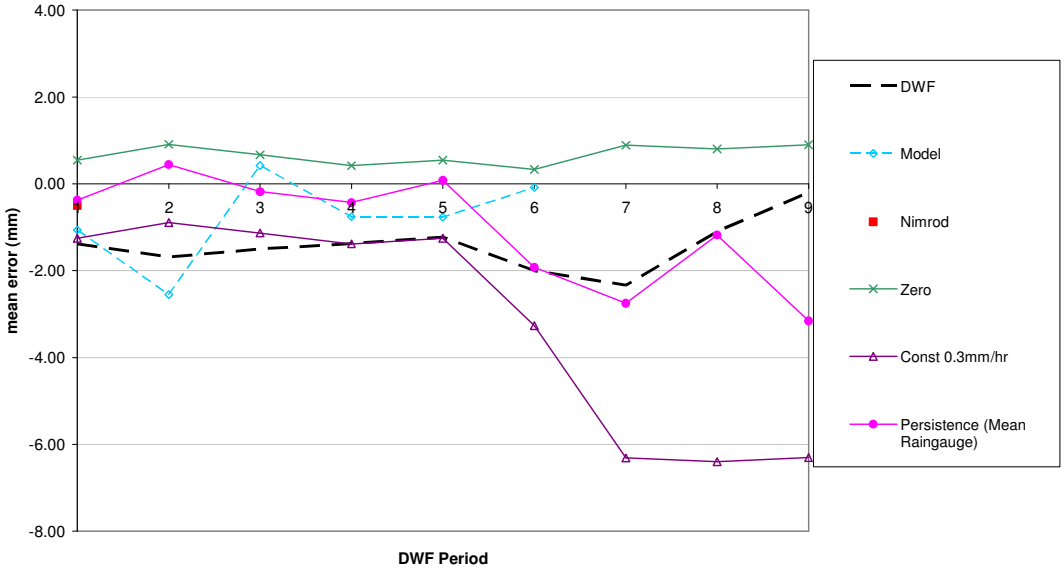


Figure 5.3.2.5 Bias measures of Daily Weather Forecast "Typical Rainfall" and comparative forecasts, Thames Northeast Area, obtained using radar areal average and mode raingauge ground truths for two case study events (12 forecast occasions).

(c) **Mean error using modal raingauge ground truth**



(d) Mean error of log rainfall using modal raingauge ground truth

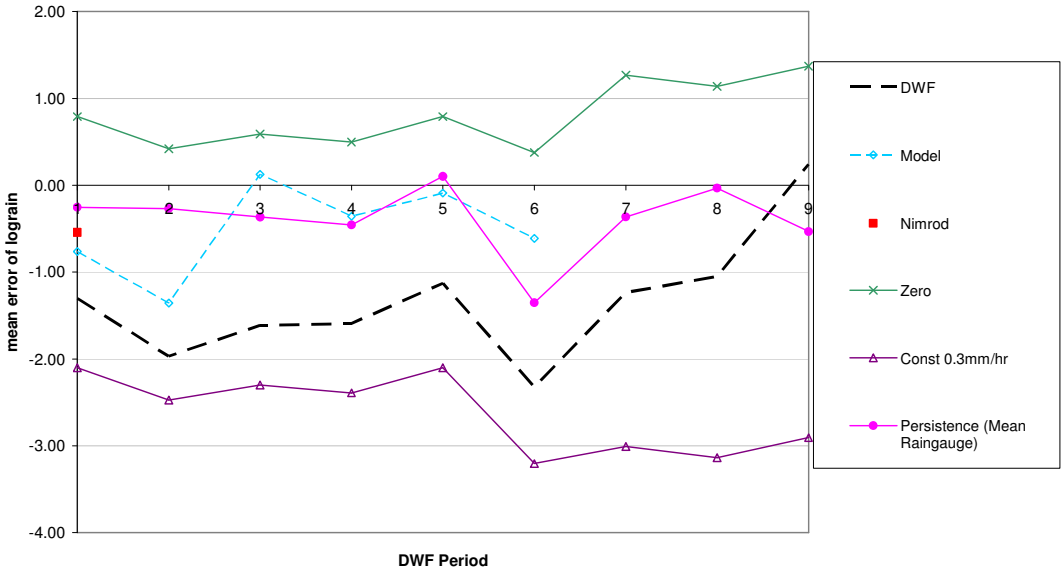


Figure 5.3.2.5 cont' Bias measures of Daily Weather Forecast "Typical Rainfall" and comparative forecasts, Thames Northeast Area, obtained using radar areal average and mode raingauge ground truths for two case study events (12 forecast occasions).

5.3.2.6 Skill Scores

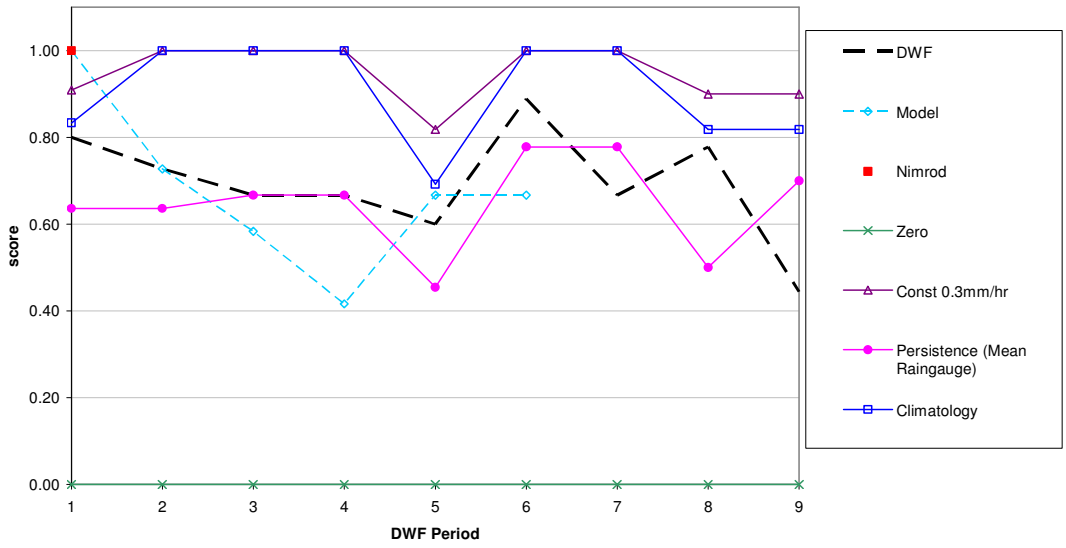
Figures 5.3.2.6 show six categorical skill scores for the "Typical" and "Max" rainfall quantities. Thresholds of 0 mm and 4 mm were found to be the most useful for this number of assessment occasions.

Figures 5.3.2.6 (a) to (e) and (h) to (k) show simple skill scores which assess the absolute performance of the forecasts. The scores shown are Critical Success Index (CSI), False Alarm Rate (FAR) and Probability of Detection (POD). CSI with a threshold of 0 mm measures correct forecasts of rain when rain occurred and additionally penalises false alarms of forecast of rain when no rain occurred. CSI with a threshold of 4mm similarly measures correct forecasts or false alarms above this amount. POD measures correct forecast of events above the threshold. FAR measures false alarms when forecasts were above the threshold but observations were below the threshold.

Figures 5.3.2.6 (f) to (h) and (m) to (o) show more complex skill scores in which the forecast performance is measures relative to random forecasts, shown as "Climatology" on the graphs. These indicate forecasts generated randomly but with the same number of forecasts exceeding the threshold as found in the observations.

For simplicity a constant threshold of 4mm has been used here although in practice it may be preferable to use a threshold dependent on the length of the forecast period.

(a) CSI for Typical Rainfall with threshold = 0 mm, areal radar ground truth



(b) CSI for Typical Rainfall with threshold = 4 mm, areal radar ground truth

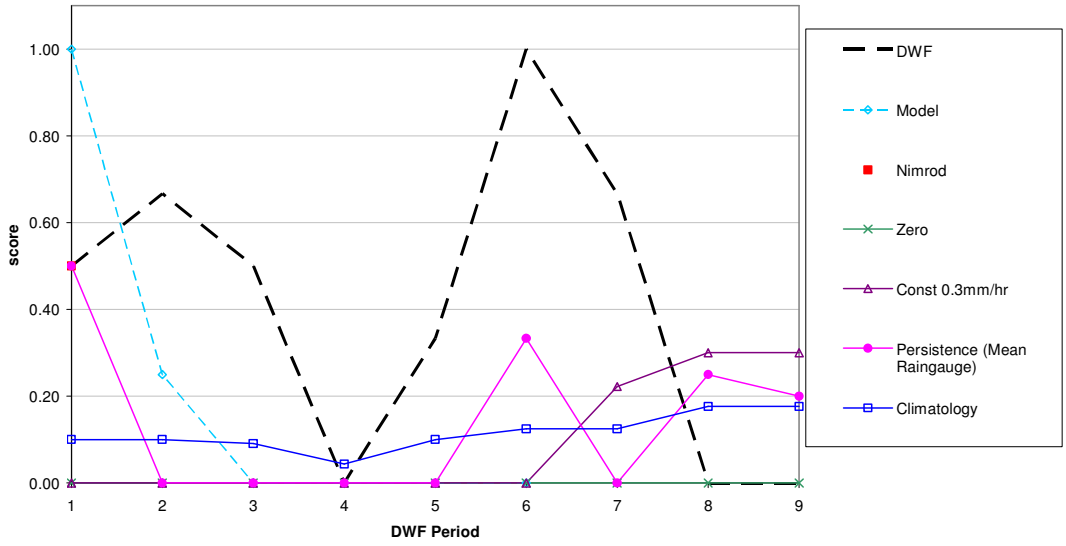
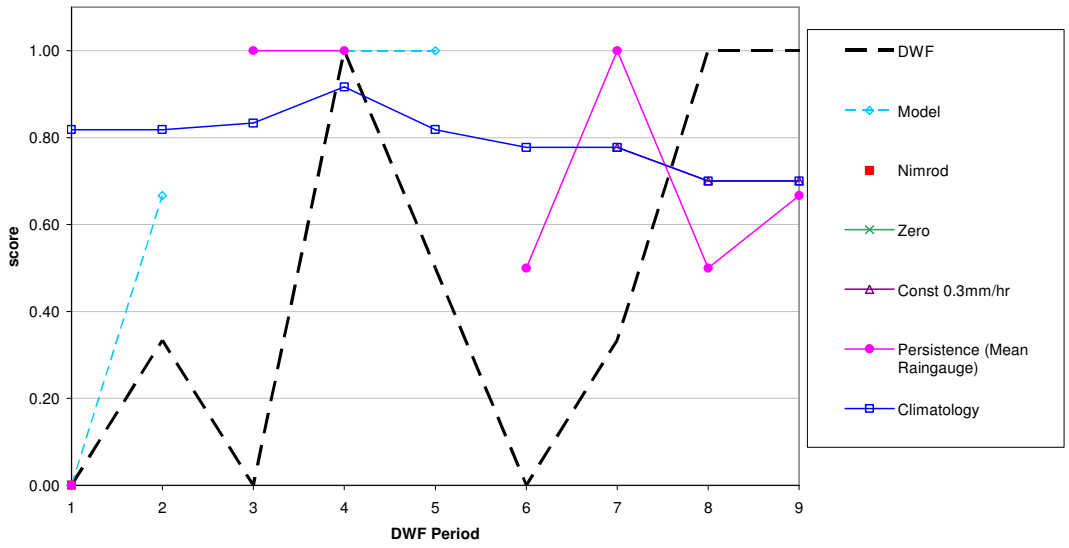


Figure 5.3.2.6 Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

(c) FAR for Typical Rainfall, with threshold = 4 mm, areal radar ground truth



(d) POD for Typical Rainfall, with threshold = 0 mm, areal radar ground truth

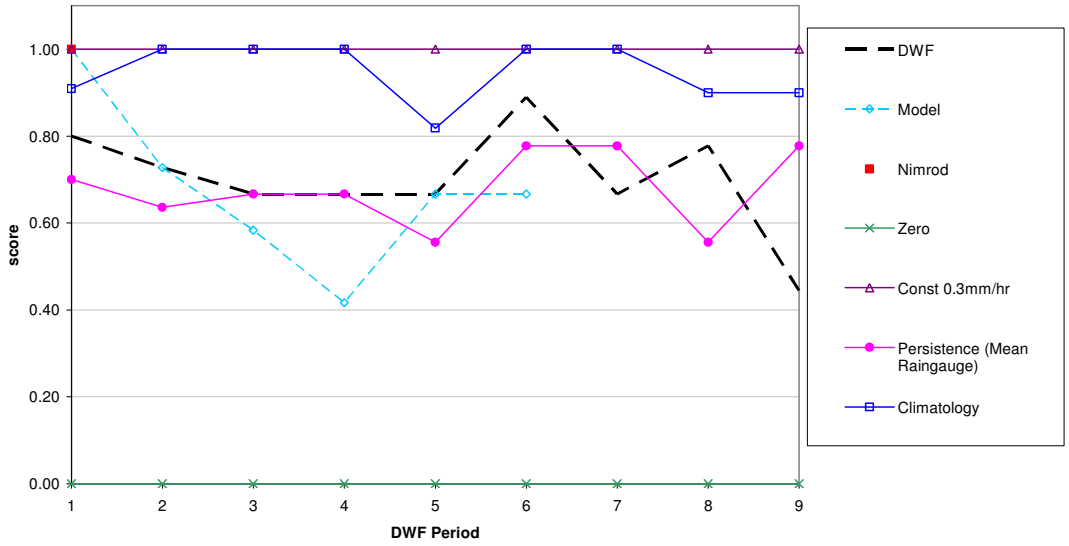
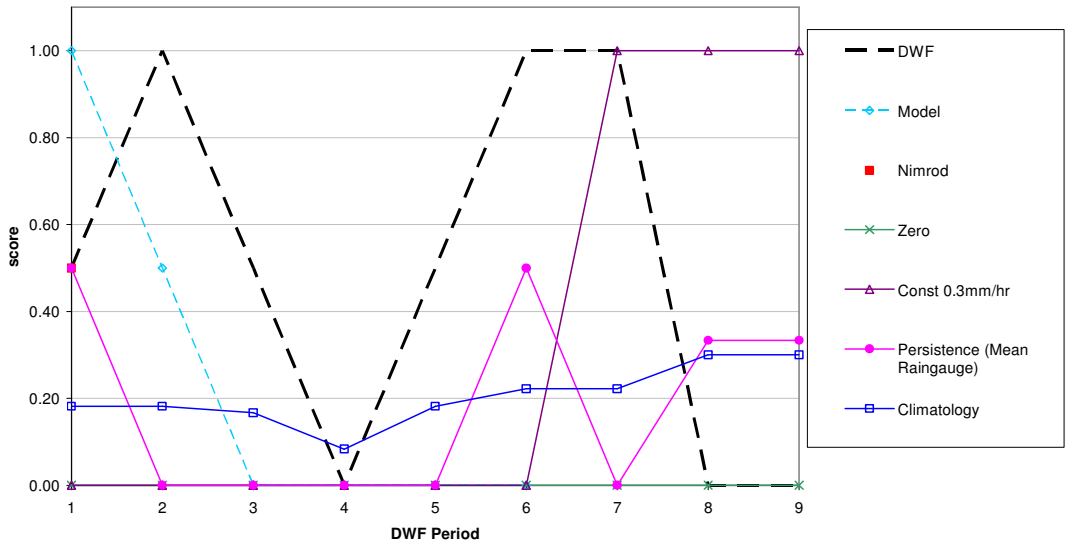


Figure 5.3.2.6 cont' Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

(e) POD for Typical Rainfall, with threshold = 4 mm, areal radar ground truth



(f) ETS for Typical Rainfall, with threshold = 4 mm, areal radar ground truth

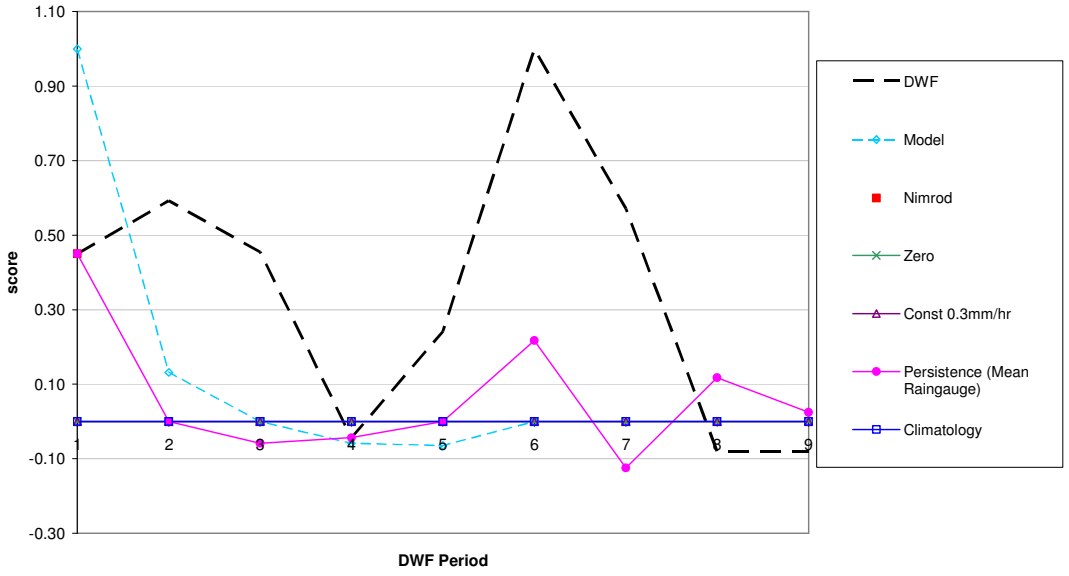
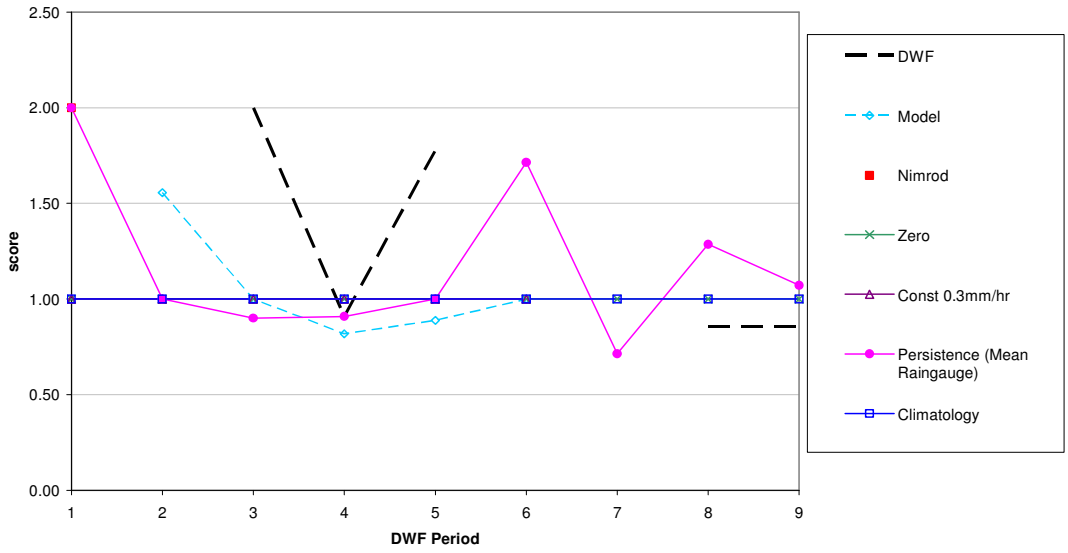


Figure 5.3.2.6 cont' Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

(g) LR1 for Typical Rainfall, with threshold = 4 mm, areal radar ground truth



(h) LR2 for Typical Rainfall, with threshold = 4 mm, areal radar ground truth

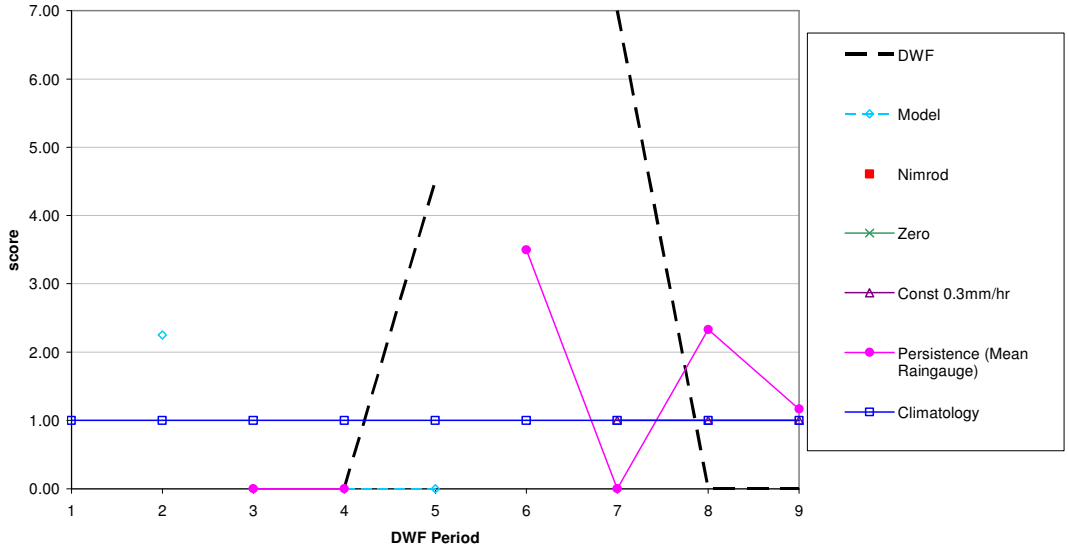
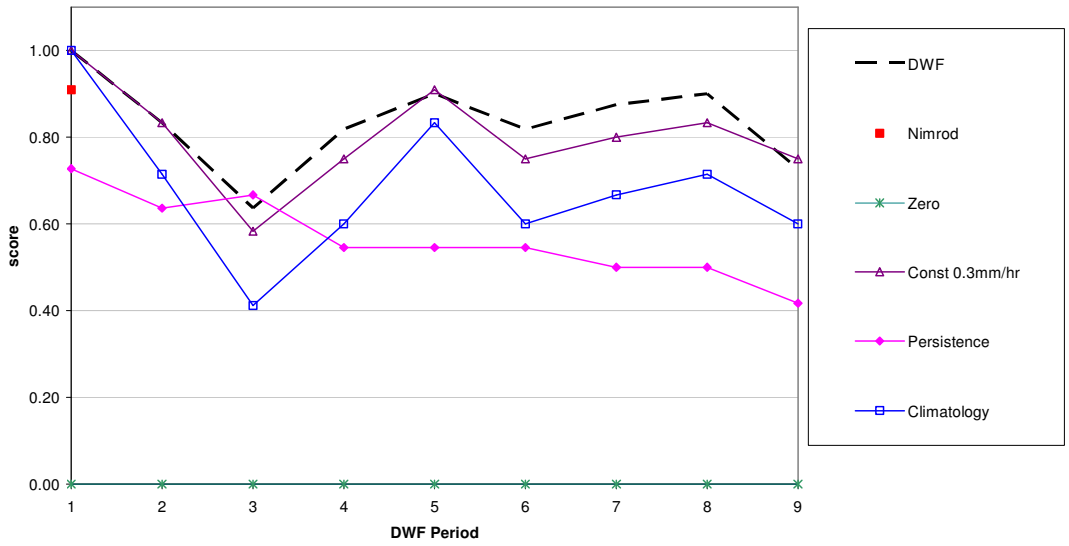


Figure 5.3.2.6 cont' Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

(i) CSI for Max Rainfall, theshold = 0 mm, raingauge ground truth



(j) CSI for Max Rainfall, threshold = 4 mm, raingauge ground truth

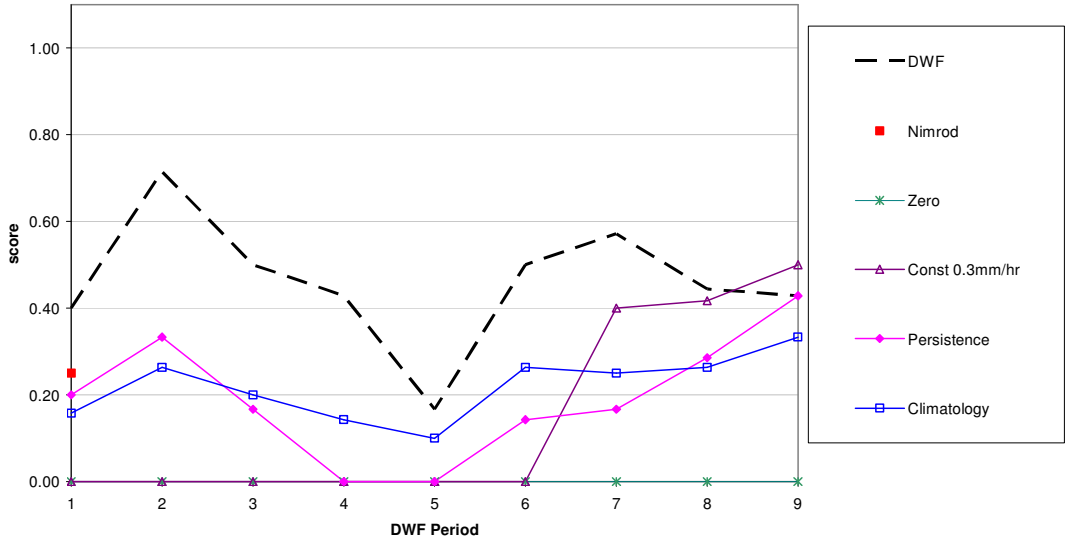
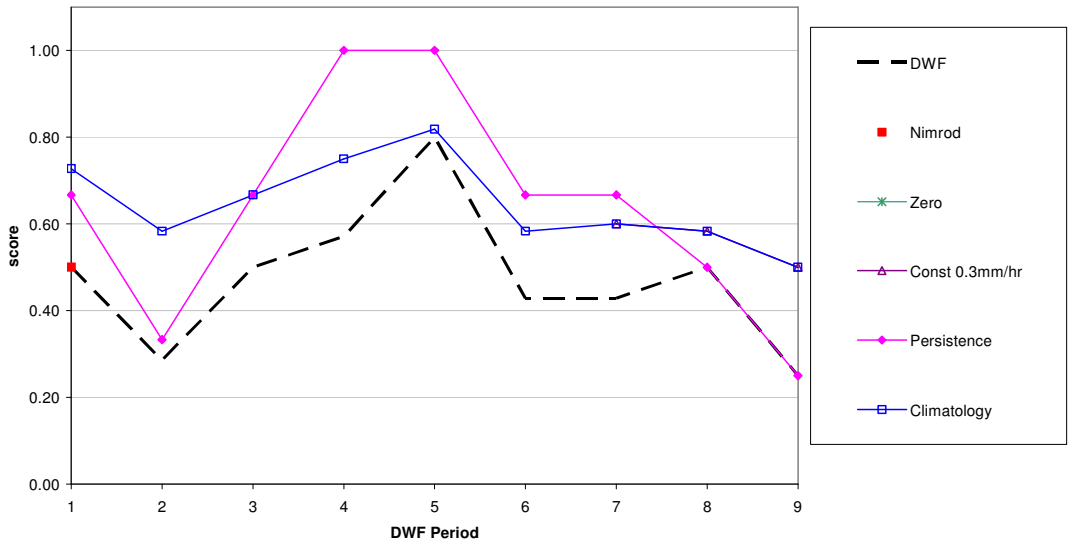


Figure 5.3.2.6 cont' Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

(k) FAR for Max Rainfall, threshold = 0 mm, raingauge ground truth



(l) POD for Max Rainfall, threshold = 0 mm, raingauge ground truth

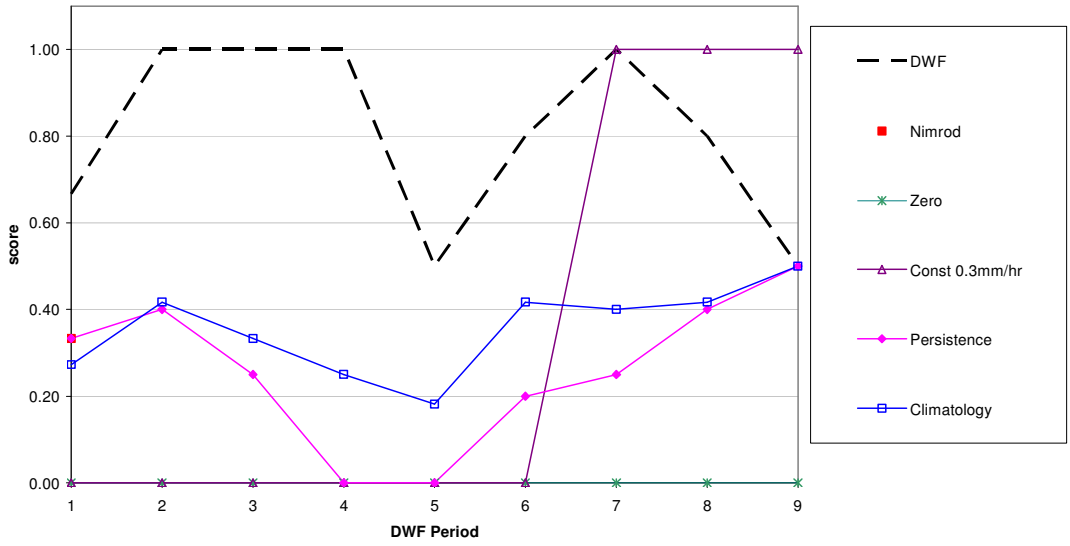
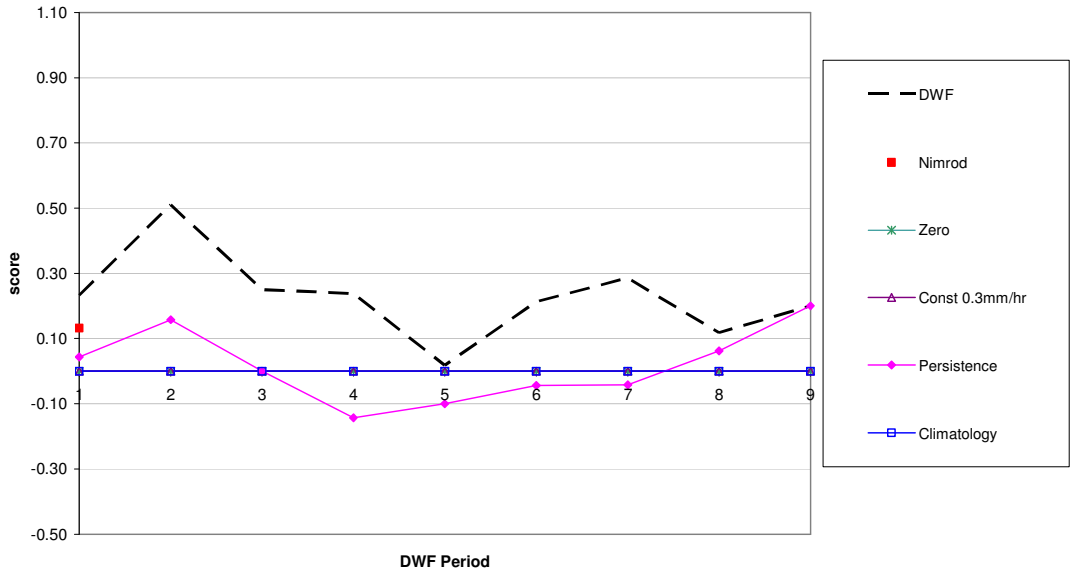


Figure 5.3.2.6 cont' Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

(m) KSS for Max Rainfall, threshold = 0 mm, raingauge ground truth



n) LR1 for Max Rainfall, threshold = 0 mm, raingauge ground truth

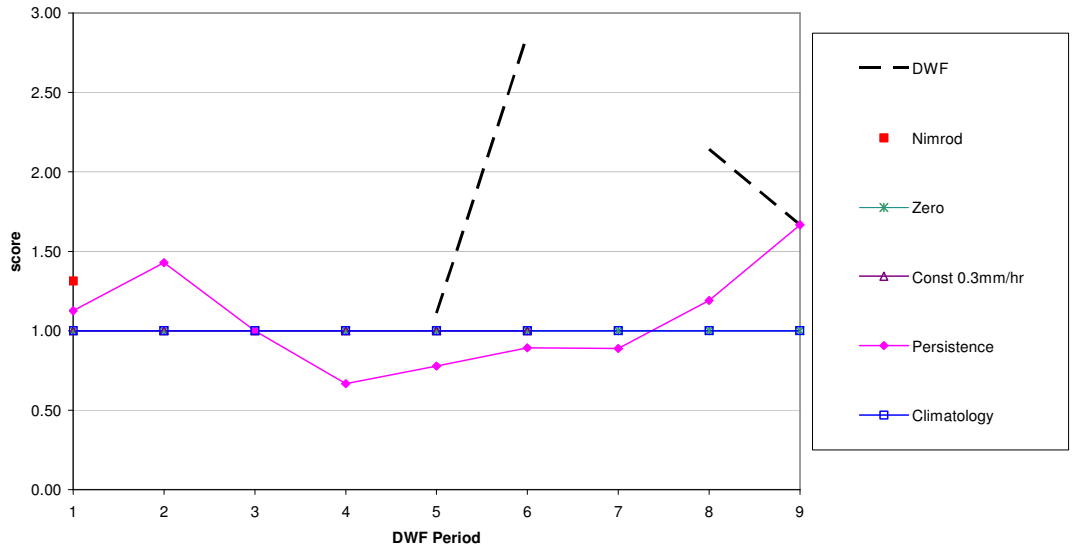


Figure 5.3.2.6 cont' Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

(o) LR2 for Max Rainfall, threshold = 0 mm, raingauge ground truth

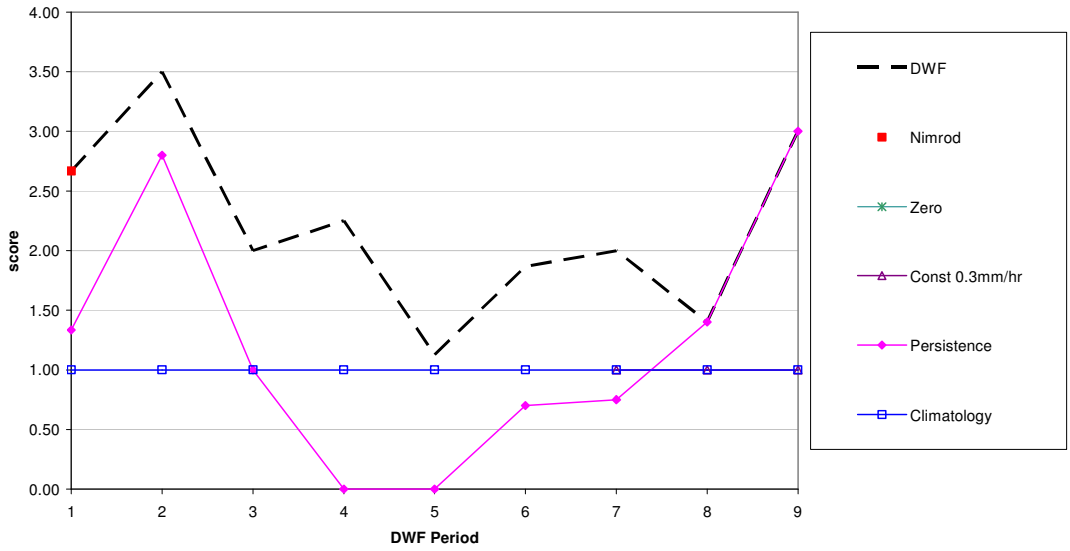


Figure 5.3.2.6 cont' Skill Scores for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, together with comparative forecasts, Thames Northeast Area, obtained for two case study events (12 forecast occasions).

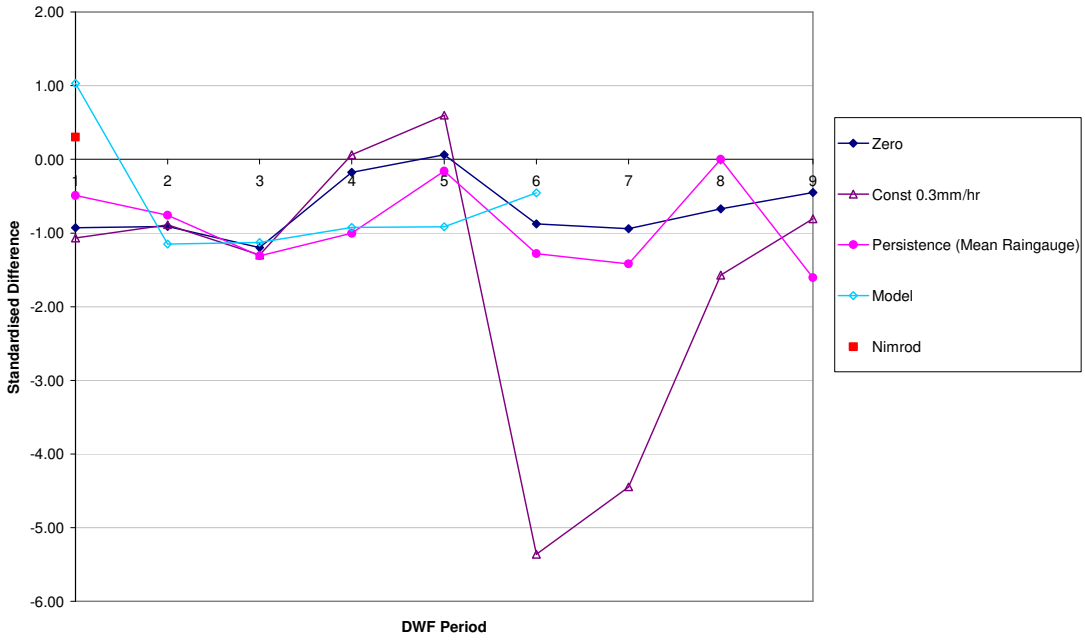
5.3.2.7 Comparison of Forecasts

Figures 5.3.2.7 (a) to (f) show the standardised differences of the root mean square error and root mean square error of log rainfall for comparative forecasts and the Daily Weather Forecast "Typical" and "Max" rainfall forecasts. In this assessment each forecast in turn is used as a "base forecast" and compared to the Daily Weather Forecast. Positive values in the graphs suggest that the base forecast in question is better than the Daily Weather Forecast. Values greater than 2.5 would indicate that there is reasonably strong evidence that the forecast is better than the Daily Weather Forecast. Similarly negative values indicate the performance is worse than the Daily Weather forecast, with values less than -2.5 indicating fairly strong evidence of this. The criterion value “2.5” is used, as discussed in Section 4.3.3, because of the small sample size.

As expected, in agreement with the raw performance measures presented in Section 5.3.2.4, the normal versions of the performance measures indicate that the Daily Weather Forecast "Typical" rainfall forecast is better than that of the Mesoscale Model, while the log versions imply the opposite. However, the evidence for these conclusions is very weak. For maximum rainfall, there is weak evidence from both measures that the Daily Weather Forecast is better than the naive forecasts. All the figures show there is some evidence that the Nimrod forecast for the first period is better than the Daily Weather Forecast.

The results here show how the use of the standardised difference, in conjunction with the usual performance measures, provides useful information about how much evidence there is that one forecast performed better than another.

(a) Standardised difference of root mean square error for Typical Rainfall, radar areal average ground truth. Positive values indicate forecast better than Daily Weather Forecast.



(b) Standardised difference of root mean square error of log rain for Typical Rainfall, radar areal average ground truth. Positive values indicate forecast better than Daily Weather Forecast.

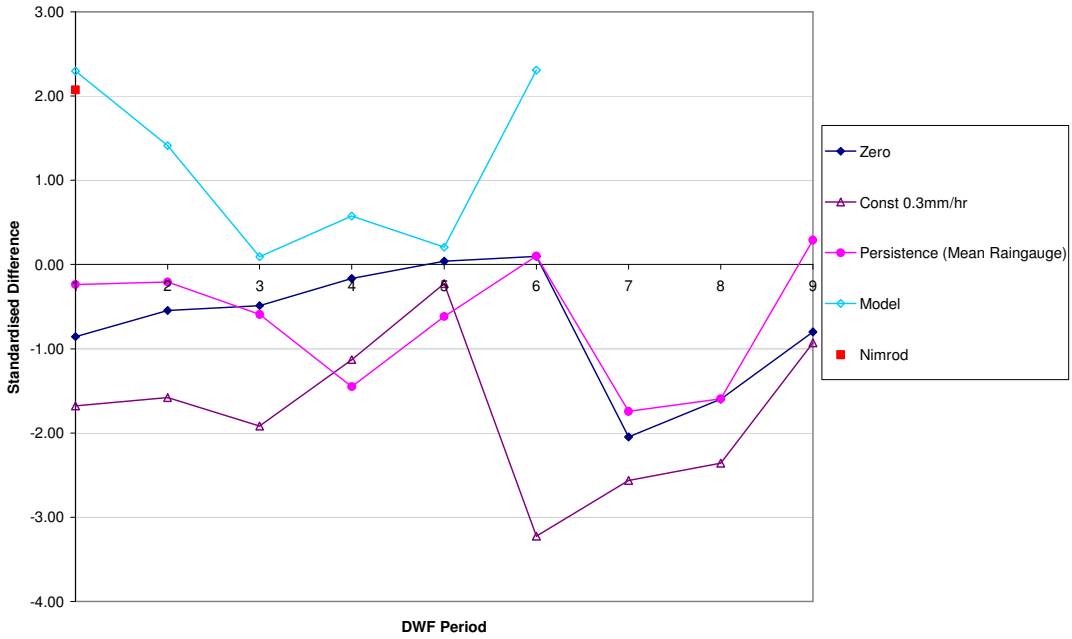
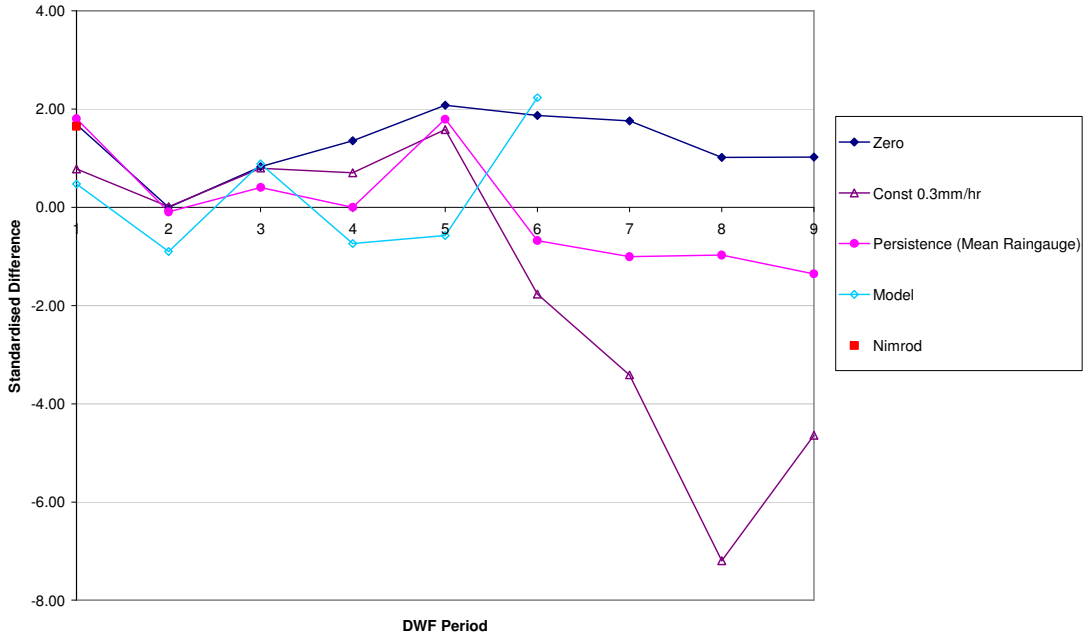


Figure 5.3.2.7 Standardised Differences of root mean square error and root mean square error of log rainfall for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, against comparative forecasts. Results shown for Thames Northeast Area, obtained using radar areal average and modal raingauge ground truths for two case study events (12 forecast occasions).

(c) Standardised difference of root mean square error of log rain for Typical Rainfall, modal raingauge ground truth. Positive values indicate forecast better than Daily Weather Forecast.



(d) Standardised difference of root mean square error for Typical Rainfall, modal raingauge ground truth. Positive values indicate forecast better than Daily Weather Forecast.

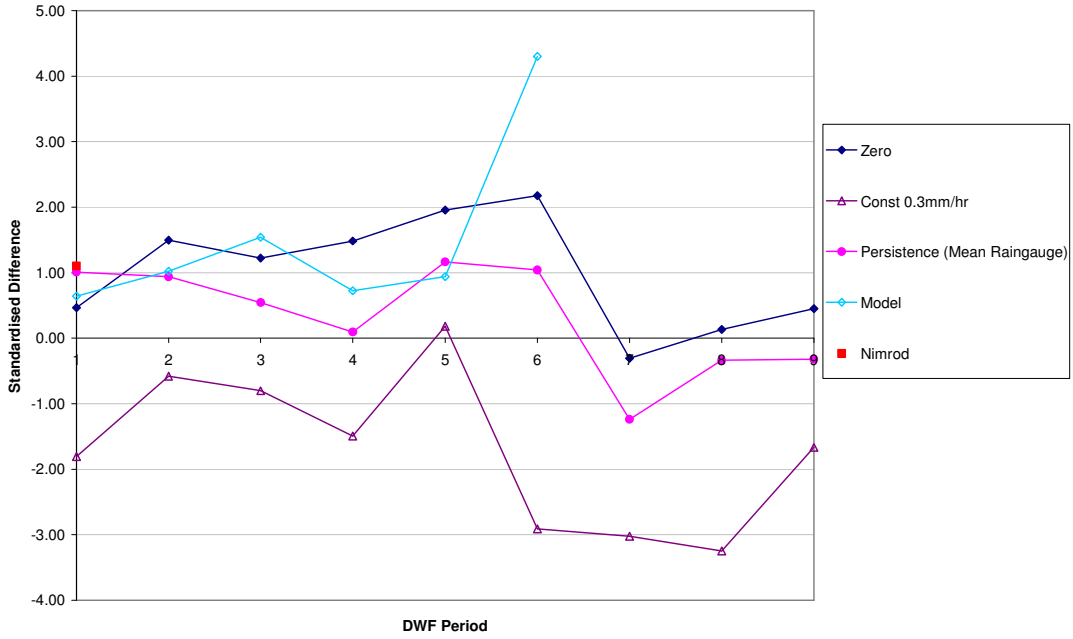
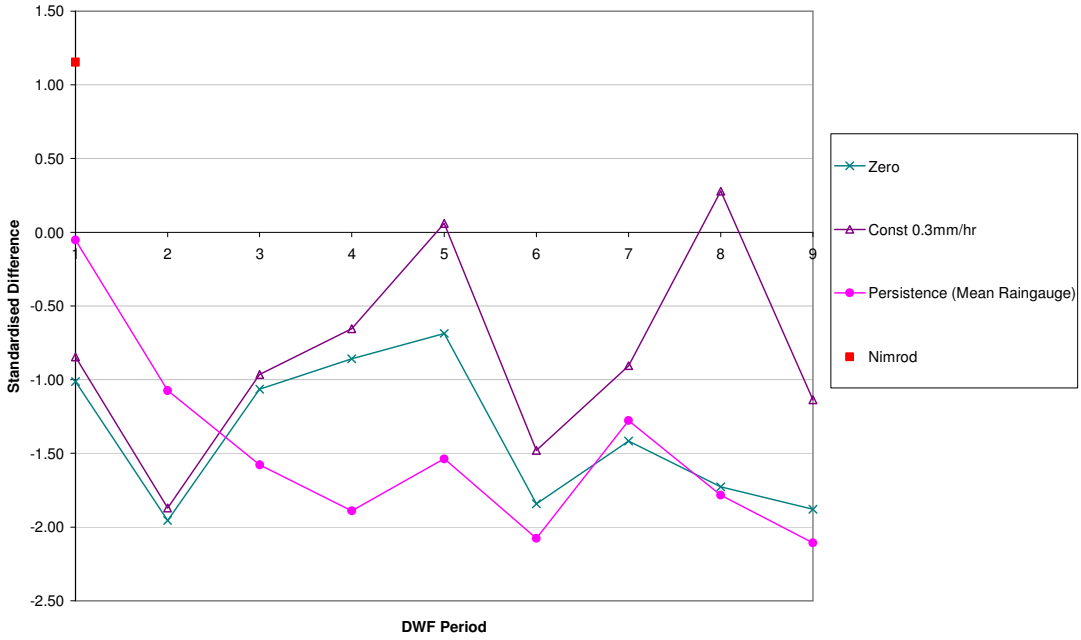


Figure 5.3.2.7 cont' Standardised Differences of root mean square error and root mean square error of log rainfall for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, against comparative forecasts. Results shown for Thames Northeast Area, obtained using radar areal average and modal raingauge ground truths for two case study events (12 forecast occasions).

(e) Root mean square error for Max Rainfall, raingauge ground truth. Positive values indicate forecast better than Daily Weather Forecast.



(f) Root mean square error of log rain for Max Rainfall, raingauge ground truth. Positive values indicate forecast better than Daily Weather Forecast.

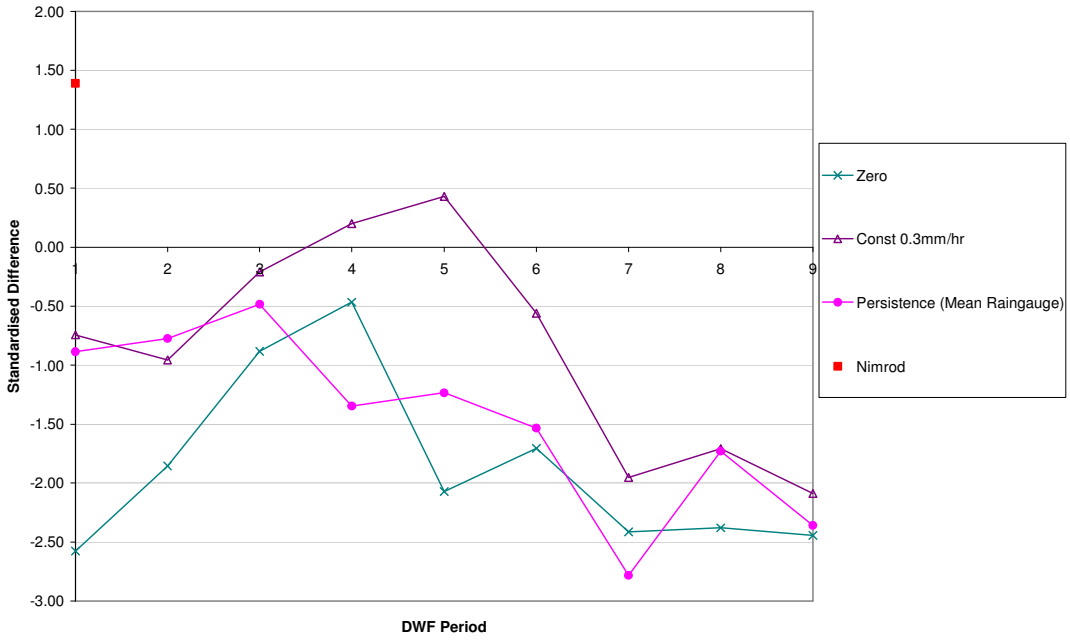


Figure 5.3.2.7 cont' Standardised Differences of root mean square error and root mean square error of log rainfall for Daily Weather Forecast "Typical Rainfall" and "Most Likely Maximum" Rainfall, against comparative forecasts. Results shown for Thames Northeast Area, obtained using radar areal average and modal raingauge ground truths for two case study events (12 forecast occasions).

5.3.3 Case Study Assessment for Northeast Region

5.3.3.1 Daily Weather Forecast Quantities

An example of a Northeast Region Daily Weather Forecast showing the relevant quantitative forecast content is given in Figure 5.3.3.1 A single section of the forecast entitled "-REGIONAL FORECAST-(RAINFALL-IN-MM)" contains the quantitative rainfall forecasts. Forecasts are given for each of the seven areas for nine periods or 6 to 24 hours over 3 days. A Met Office document giving instructions for the construction of the Daily Weather Forecast does not specify the exact nature of the forecast quantity. For this assessment it has been assumed that, as for the Northwest forecasts (also issued by Met Office Manchester), the quantity is the spatial average rainfall accumulation.

```
+
+-REGIONAL FORECAST-(RAINFALL-IN-MM)------------------------+
+                        CHVT W.PN CN.PN S.PN NE.C MOOR V.WD
+
10 Aug 02 0001-0600       02   08   01   07   02   01   01
10 Aug 02 0600-1200       04   06   05   02   06   06   02
10 Aug 02 1200-1800       02   03   03   03   02   03   03
10 Aug 02 1800-2400       00   00   00   00   00   00   00
11 Aug 02 0001-0600       00   00   00   00   00   00   00
11 Aug 02 0600-1200       00   00   00   00   00   00   00
11 Aug 02 1200-2400       05   02   02   00   05   01   01
12 Aug 02 0001-1200       01   00   00   00   01   00   00
12 Aug 02 1200-2400       00   00   00   00   00   00   00
?END
+
```

Figure 5.3.3.1 Section of Daily Weather Forecast for Northeast Region containing quantitative rainfall forecasts.

Table 5.3.3.1 summarises the forms of ground truth and comparative forecasts considered in this assessment. Although all seven Daily Weather Forecast areas in Northeast region were included in the case study analysis, in order to be concise only the results for the "North East Coast" and "South Pennines" area are presented here.

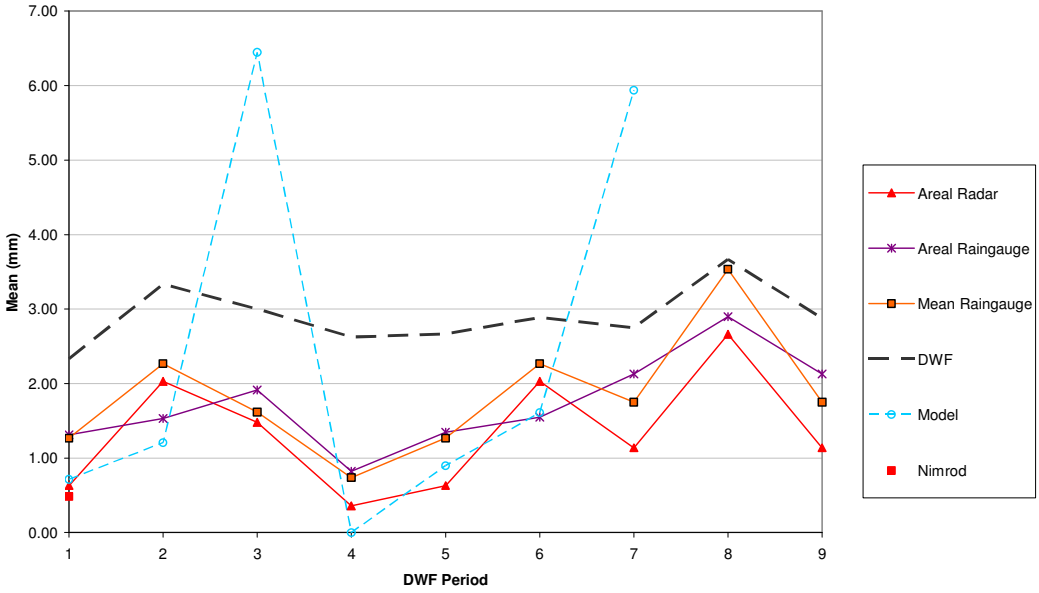
Table 5.3.3.1 Summary of target quantities, ground truths and comparative forecasts for Northeast Region Daily Weather Forecasts. Averages refer to spatial averaging carried out on raingauge and radar values which have first been accumulated over the appropriate period.

Quantity: Rainfall Accumulation (mm)	
Ground truths	**Comparative forecasts**
Raingauge • Mean • Multiquadric interpolated areal average	**Alternative forecast sources** • Nimrod forecast accumulation areal average. (Day 1 Period 1 only) • Mesoscale model areal average. (Days 1 and 2 only) • Persistence based on previous 6 hours mean raingauge accumulation.
Radar • Areal average	**Naive forecasts** • Fixed value of 0 mm. • Fixed value of 0.3 mm h^{-1} over the forecast period.

5.3.3.2 Basic Statistics of Case Study Data

Figures 5.3.3.2 (a) to (f) present basic statistics of each ground truth and forecast quantity considered in the assessment for the "North East Coast" and "South Pennines" areas. Although these statistics provide a useful reference for discussion in Sections 5.3.3.3 - 5.3.3.7 it is difficult to draw any firm conclusions about performance of the forecasts from the figures presented here. The difference between the plots for mean and median rainfall are consistent with short intense rainfall periods over the 9 days, and so it is unlikely that any conclusions about performance can be obtained by simply considering averages of very high and zero rainfall amounts over this short period. The plots are useful in considering the alternative forms of ground truth and this is discussed in Section 5.3.3.3.

(a) Mean of ground truths and forecasts across 9 assessment occasions of the case study, North East Coast Area



(b) Mean of ground truths and forecasts across 9 assessment occasions of the case study, South Pennines Area

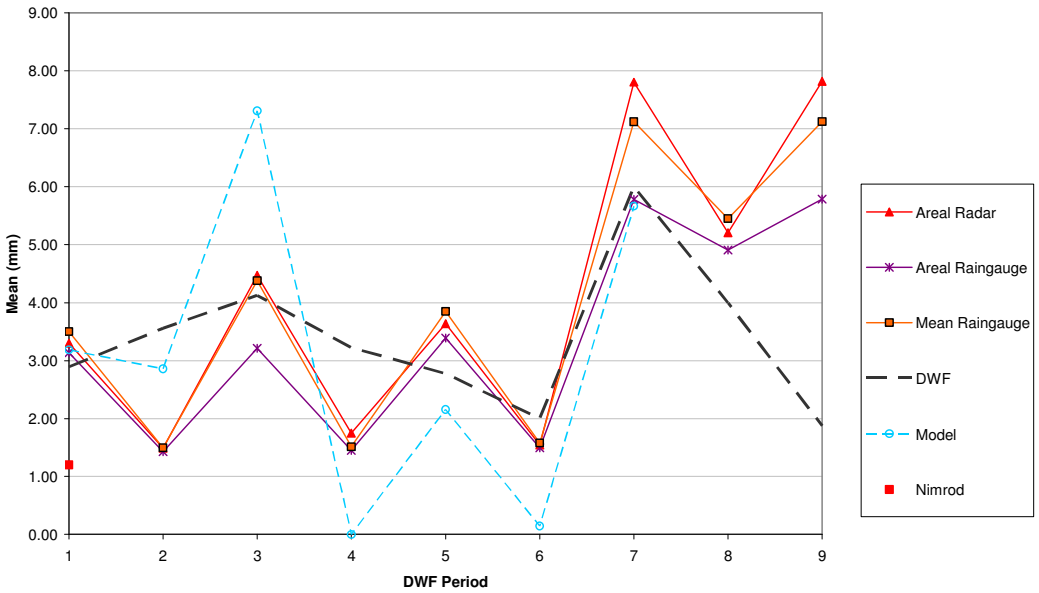
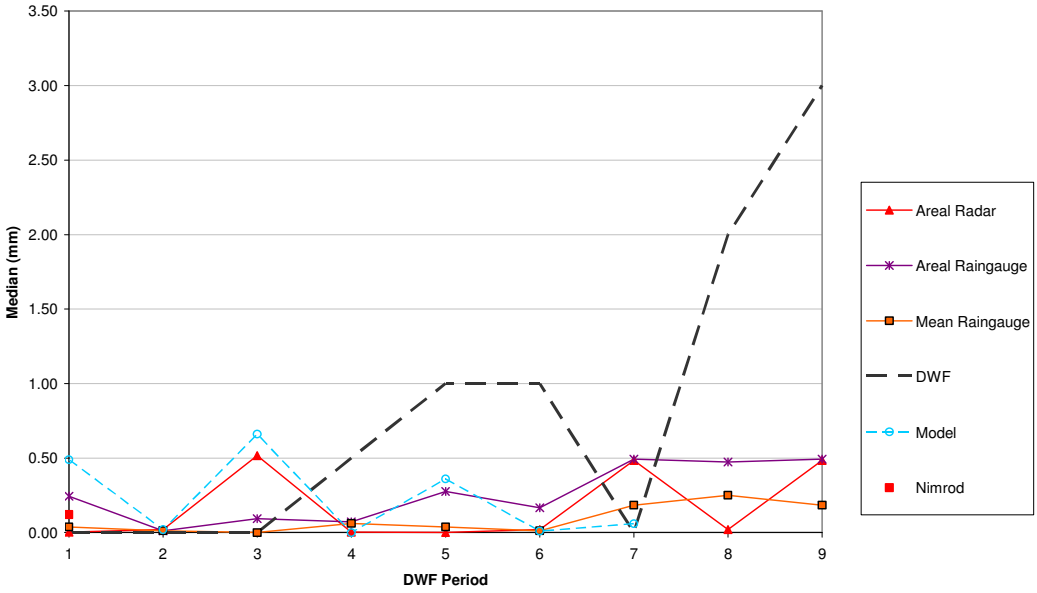


Figure 5.3.3.2 Basic statistics of datasets used for case study assessment. Northeast Region "North East Coast" and "South Pennines" areas.

(c) Median of ground truths and forecasts across 9 assessment occasions of the case study, North East Coast Area



(d) Median of ground truths and forecasts across 9 assessment occasions of the case study, South Pennines Area

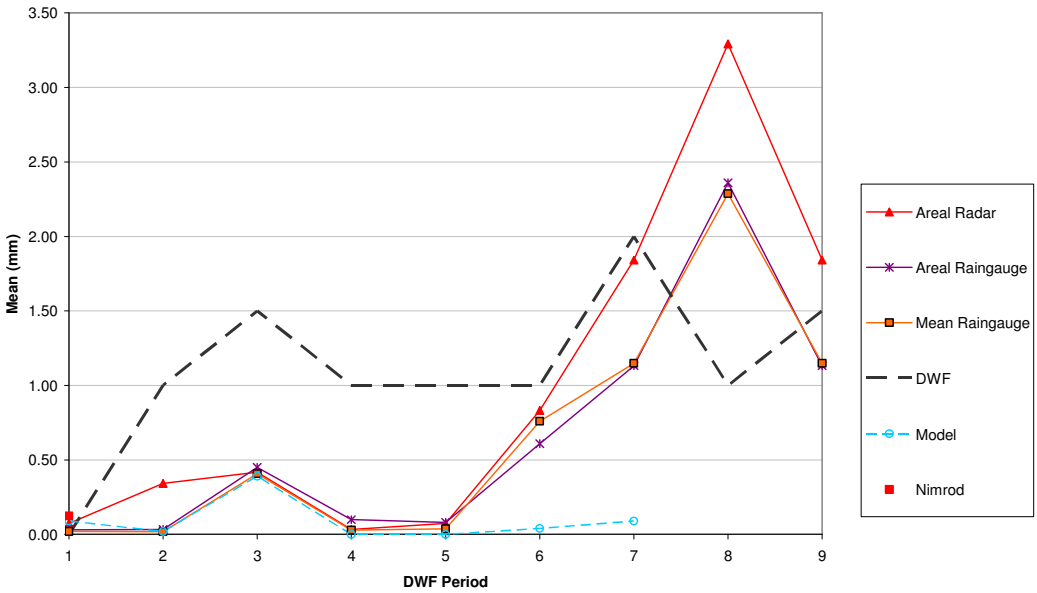
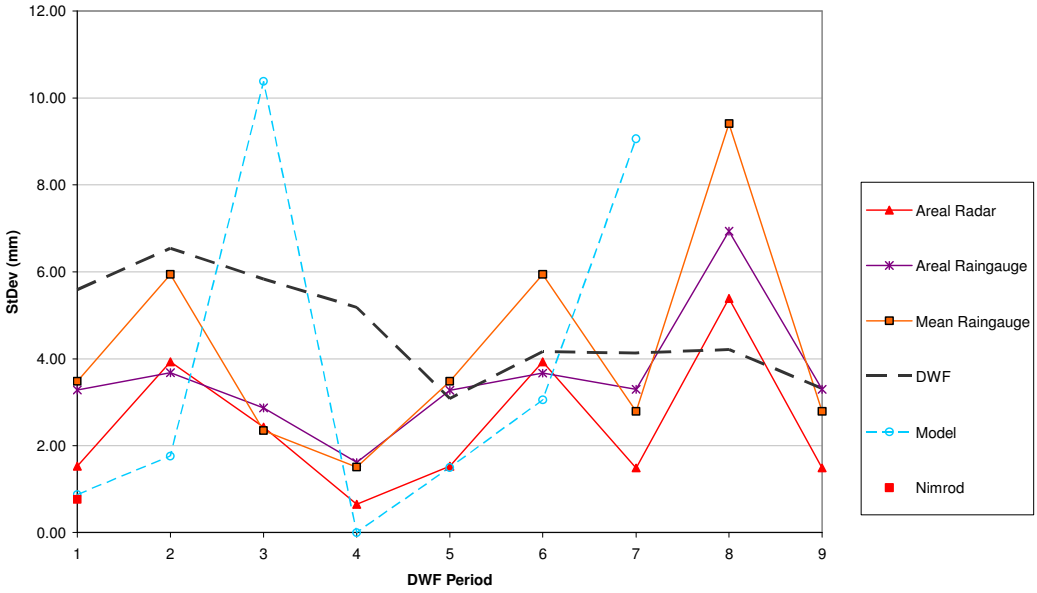


Figure 5.3.3.2 cont' Basic statistics of datasets used for case study assessment. Northeast Region "North East Coast" and "South Pennines" areas.

(e) Standard deviation of ground truths and forecasts across 9 assessment occasions of the case study, North East Coast Area



(f) Standard deviation of ground truths and forecasts across 9 assessment occasions of the case study, South Pennines Area

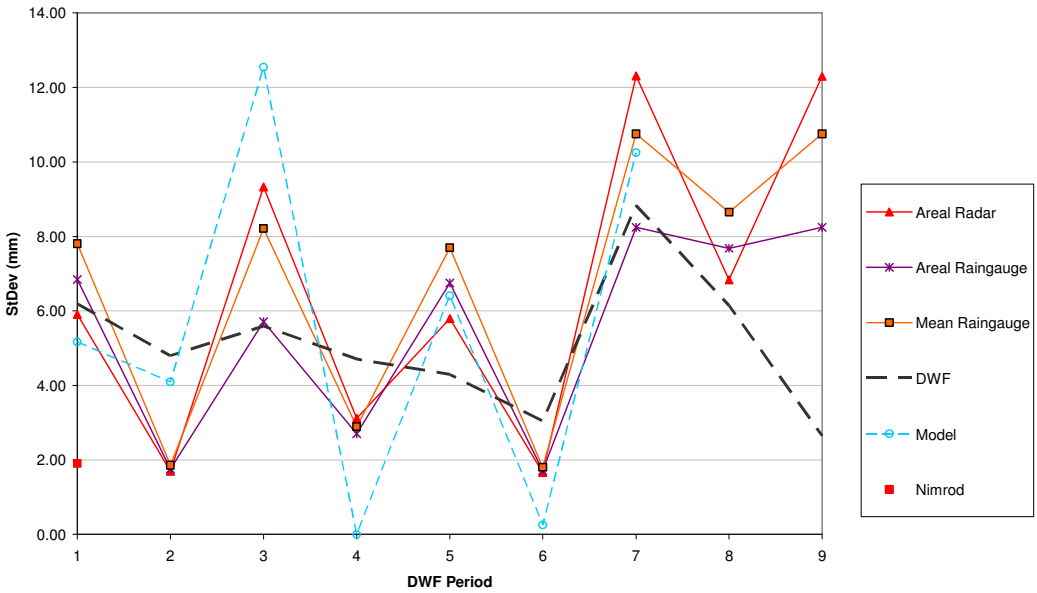


Figure 5.3.3.2 cont' Basic statistics of datasets used for case study assessment. Northeast Region "North East Coast" and "South Pennines" areas.

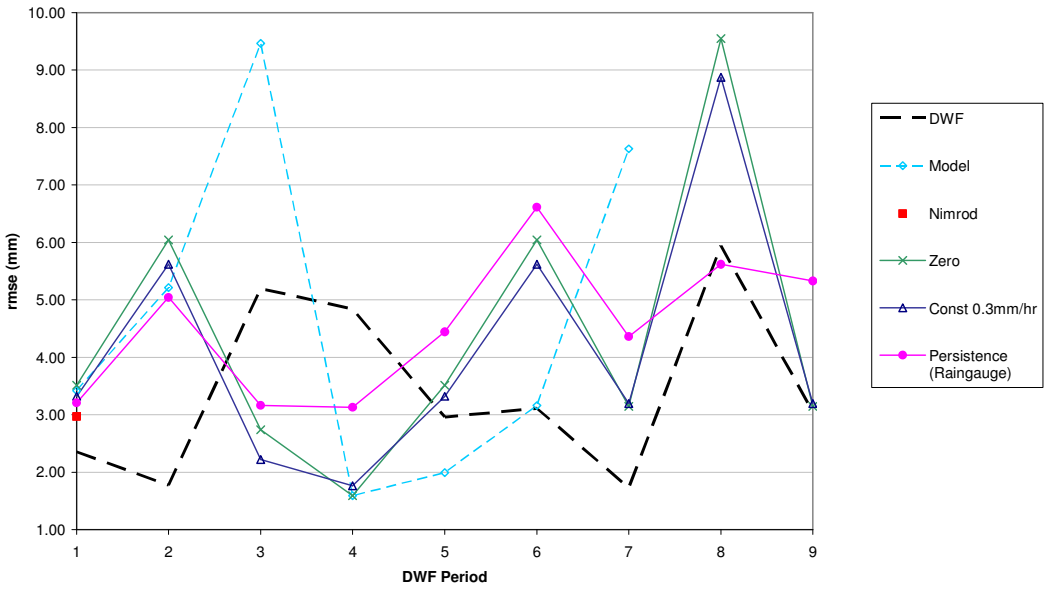
5.3.3.3 Selection of suitable forms of ground truth

The basic statistics presented in Section 5.3.3.2 indicate that the three forms of ground truth considered have similar statistical characteristics, at least over the case study event considered here. It therefore seems reasonable to proceed with the assessment using the mean raingauge as the ground truth.

5.3.3.4 Raw assessment measures

Figure 5.3.3.4 (a) to (d) present the normal and log versions of the root mean square error assessment measure for the two Daily Weather Forecast areas, using the mean raingauge ground truth.

(a) Root mean square error for North East Coast Area



(b) Root mean square error for South Pennines Area

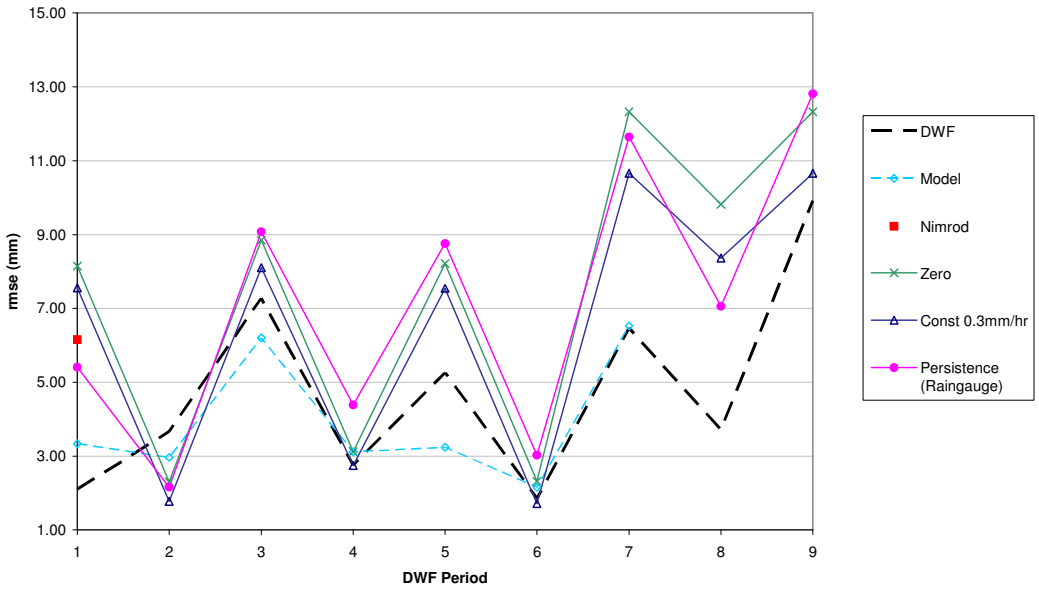
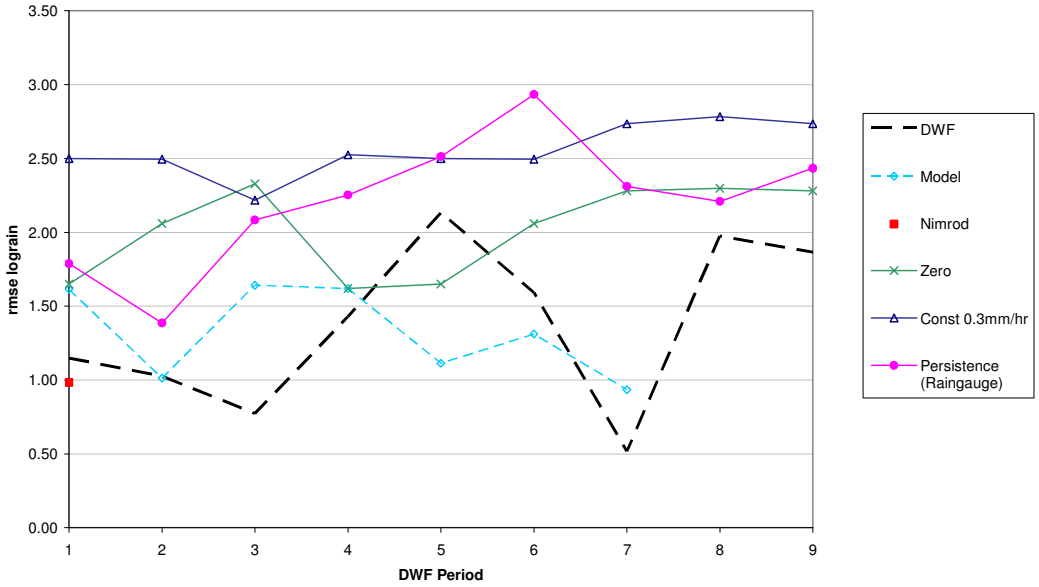


Figure 5.3.3.4 Raw performance measures for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" and "South Pennines" areas. Case study with 9 assessment occasions.

(c) Root mean square error of log rainfall for North East Coast Area



(d) Root mean square error of log rainfall for South Pennines Area

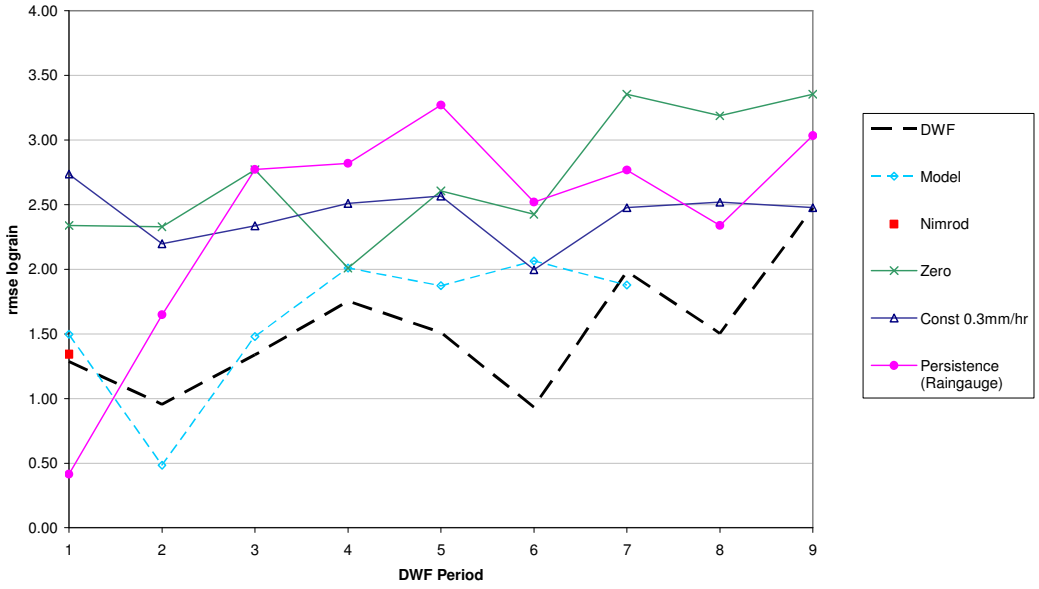
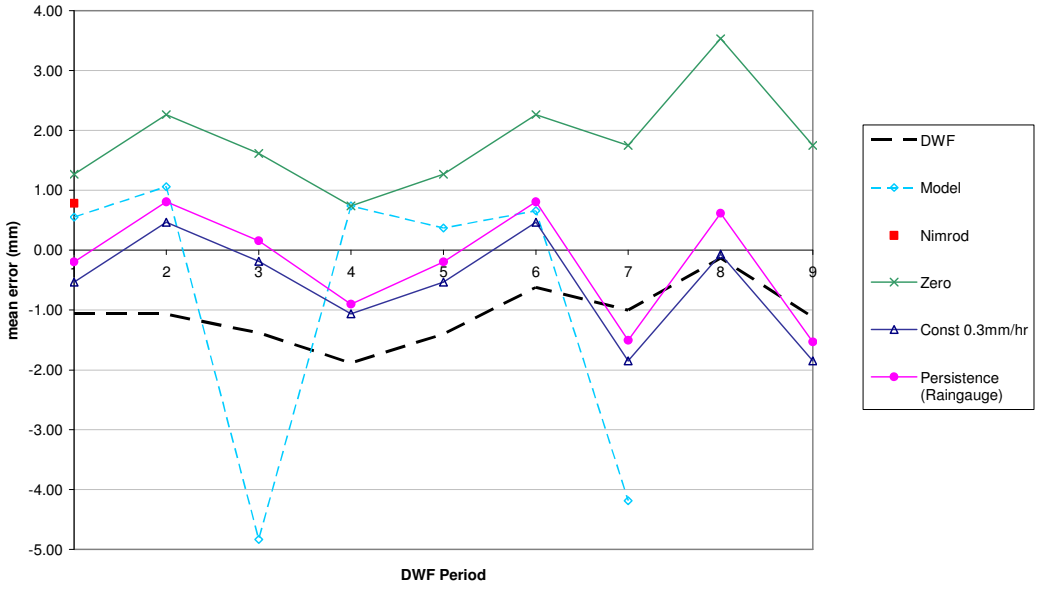


Figure 5.3.3.4 cont' Raw performance measures for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" and "South Pennines" areas. Case study with 9 assessment occasions.

5.3.3.5 Measures of Bias

Figure 5.3.3.5 presents bias measures for the Daily Weather Forecast and comparative forecasts for the two areas using the mean raingauge ground truth.

(a) Mean error for North East Coast Area



(b) Mean error for South Pennines Area

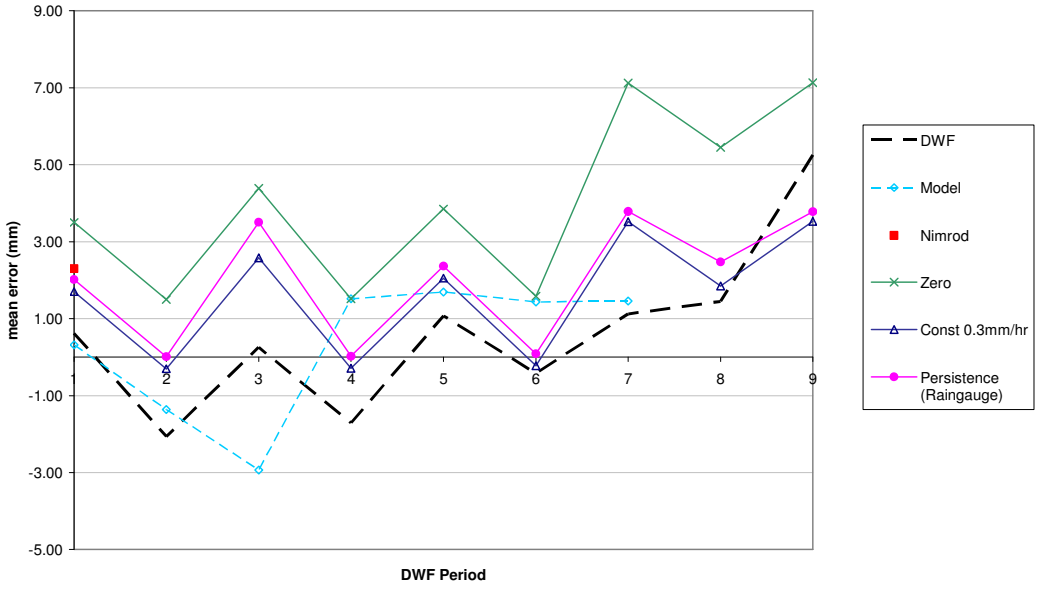
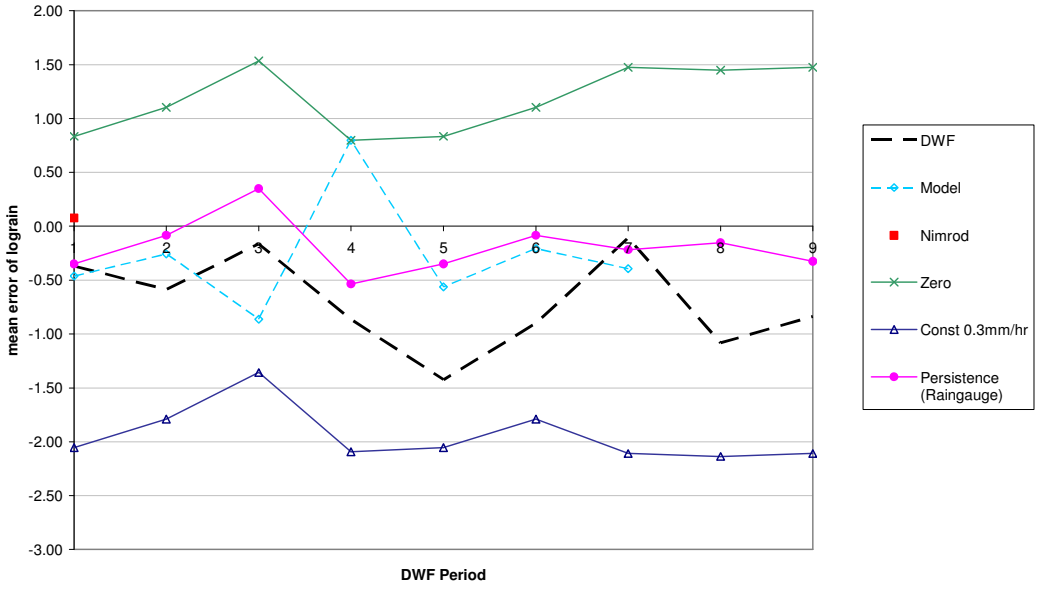


Figure 5.3.3.5 Bias measures for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" and "South Pennines" areas. Case study with 9 assessment occasions.

(c) Mean error of log rainfall for North East Coast Area



(d) Mean error of log rainfall for South Pennines Area

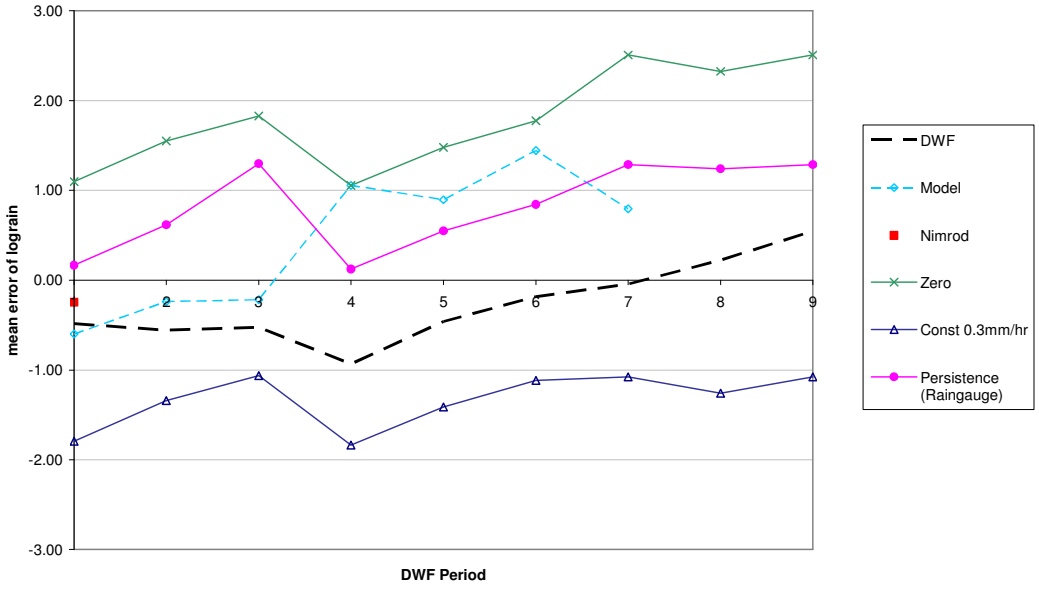
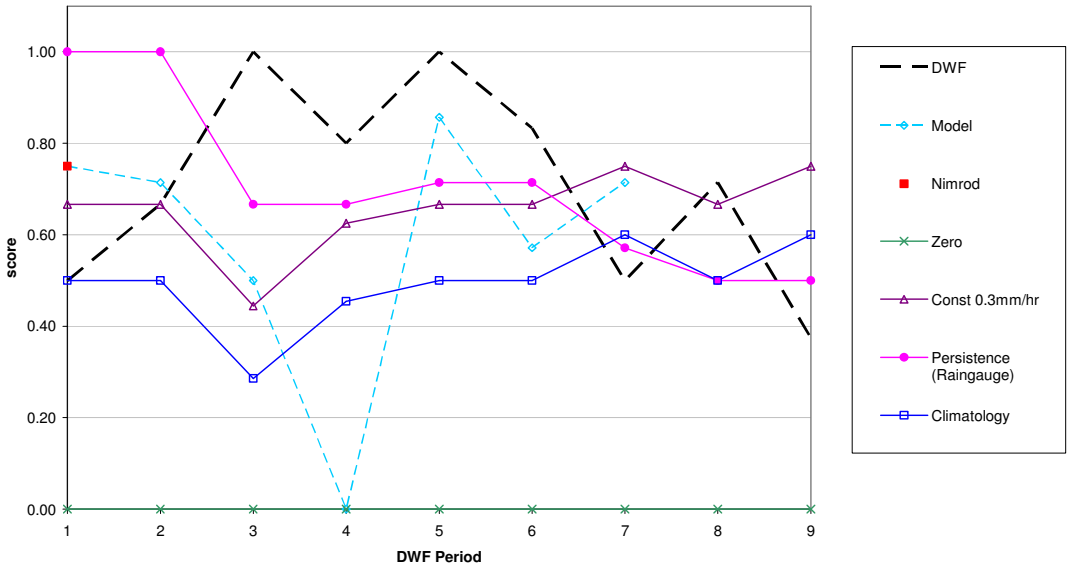


Figure 5.3.3.5 cont' Bias measures for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" and "South Pennines" areas. Case study with 9 assessment occasions.

5.3.3.6 Skill Scores

Figure 5.3.3.6 shows six category skill scores for the Daily Weather Forecasts and comparative forecasts, using the mean raingauge ground truth. In order to be concise only the result for the "North East Coast" are included.

(a) CSI threshold = 0 mm



(b) CSI threshold = 4 mm

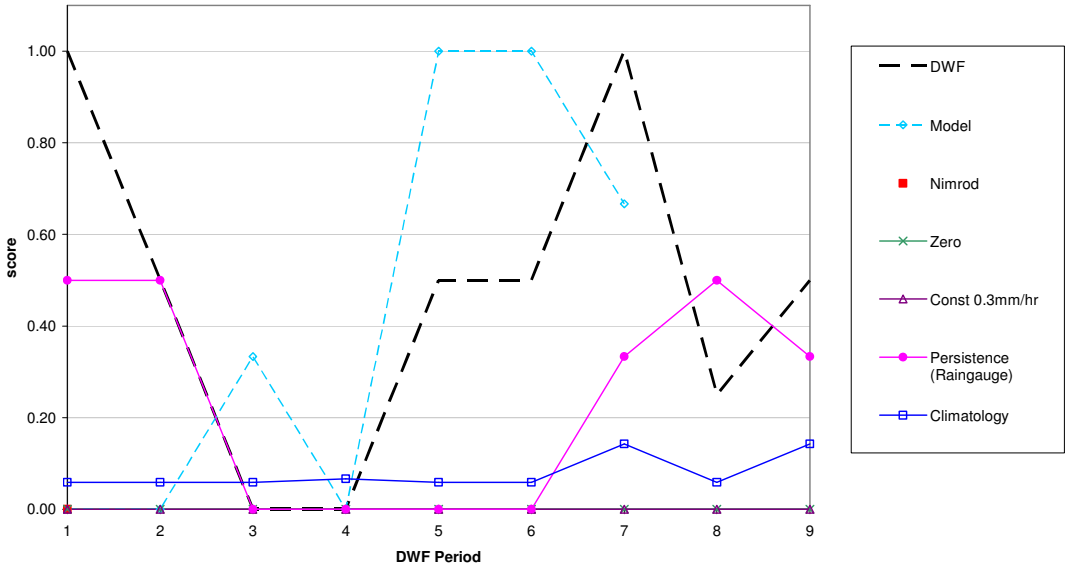
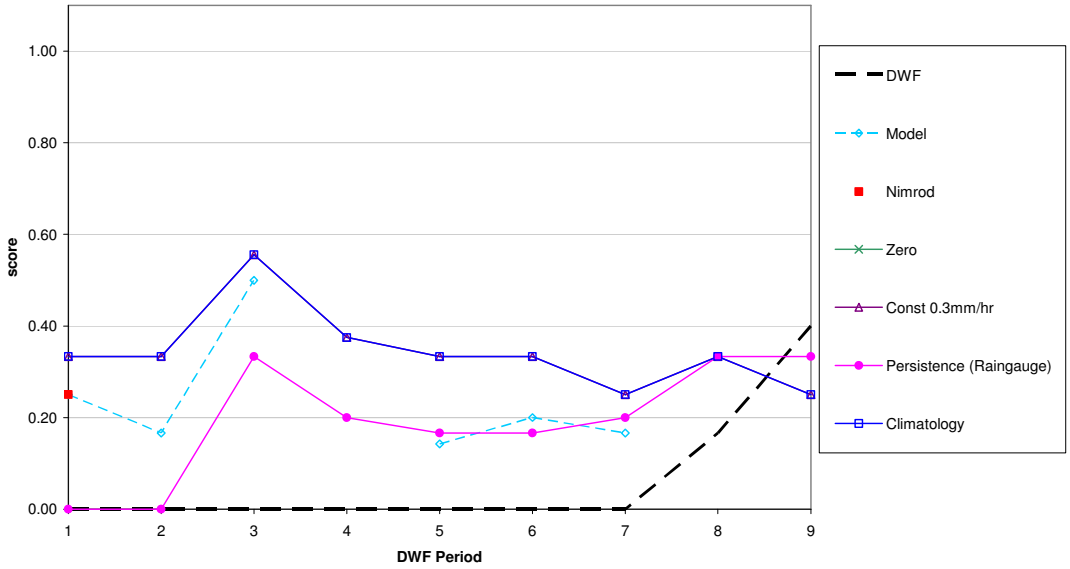


Figure 5.3.3.6 Skill scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" area. Case study with 9 assessment occasions.

(c) FAR threshold = 0 mm



(d) FAR threshold = 4 mm

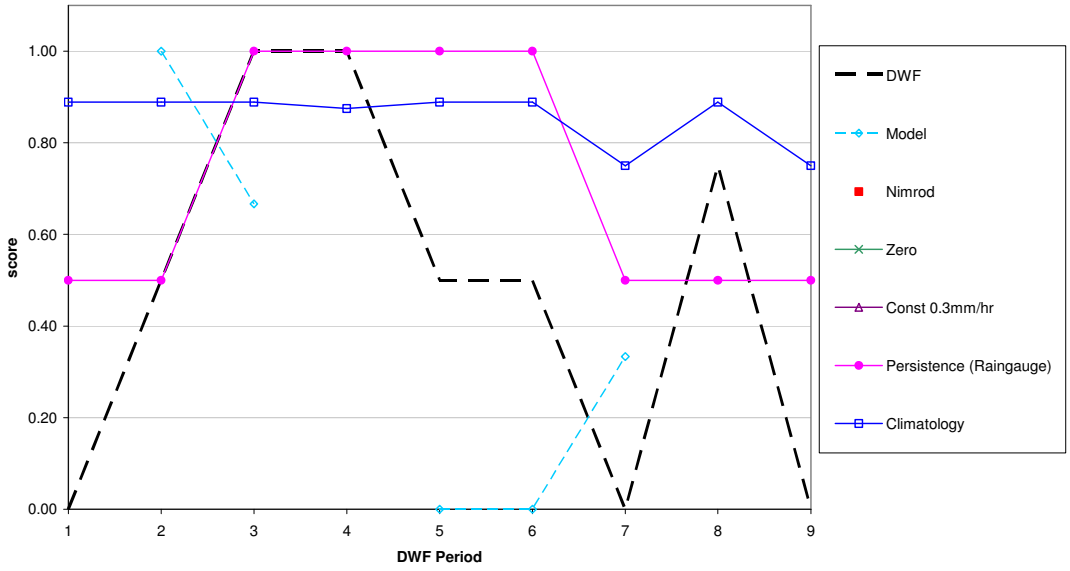
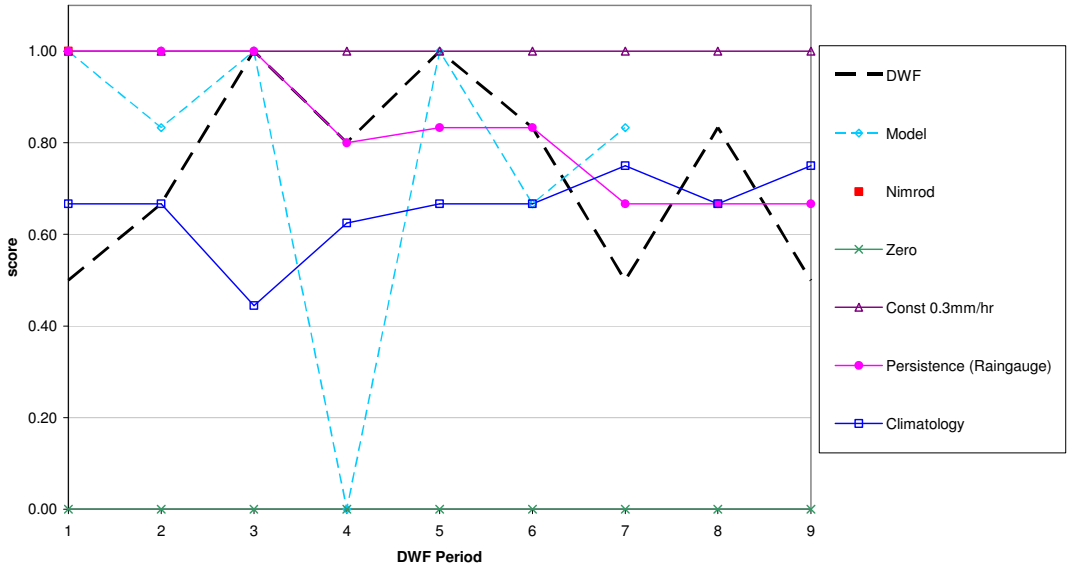


Figure 5.3.3.6 cont' Skill scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" area. Case study with 9 assessment occasions.

(e) POD threshold = 0 mm



(f) POD threshold = 4 mm

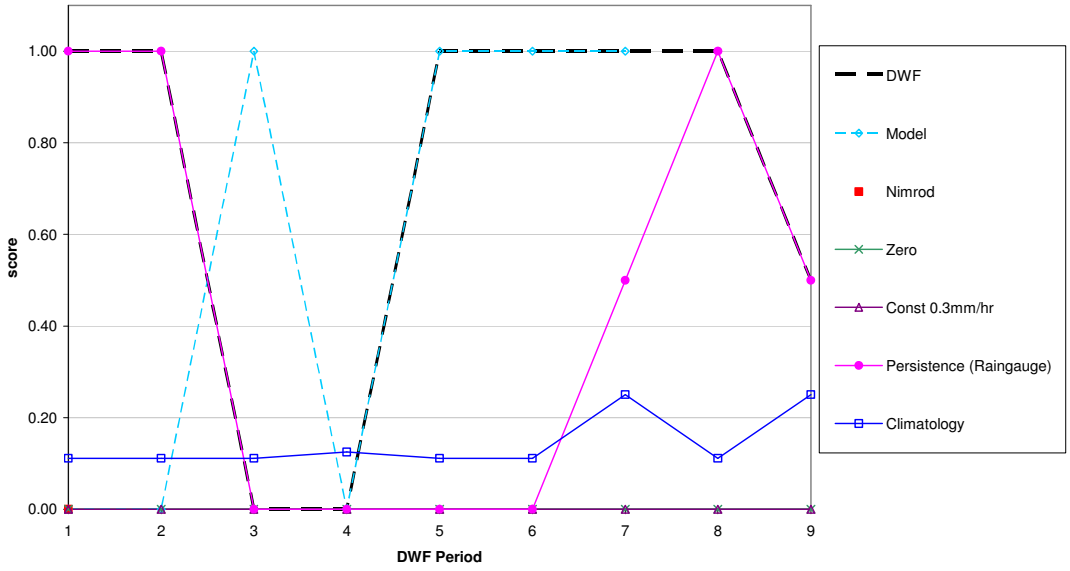
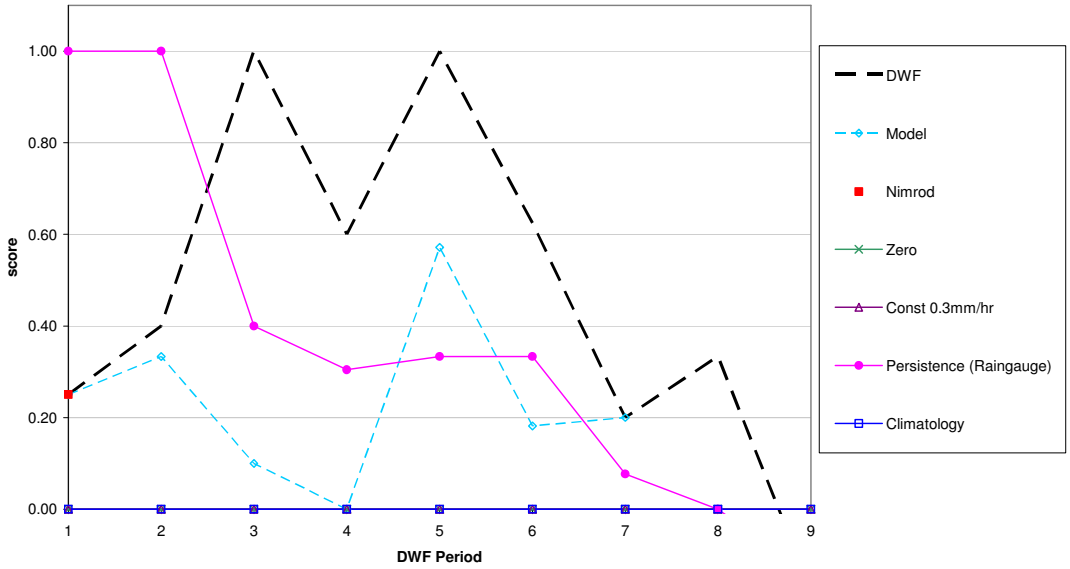


Figure 5.3.3.6 cont' Skill scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" area. Case study with 9 assessment occasions.

(g) ETS threshold = 0 mm



(h) ETS threshold = 4 mm

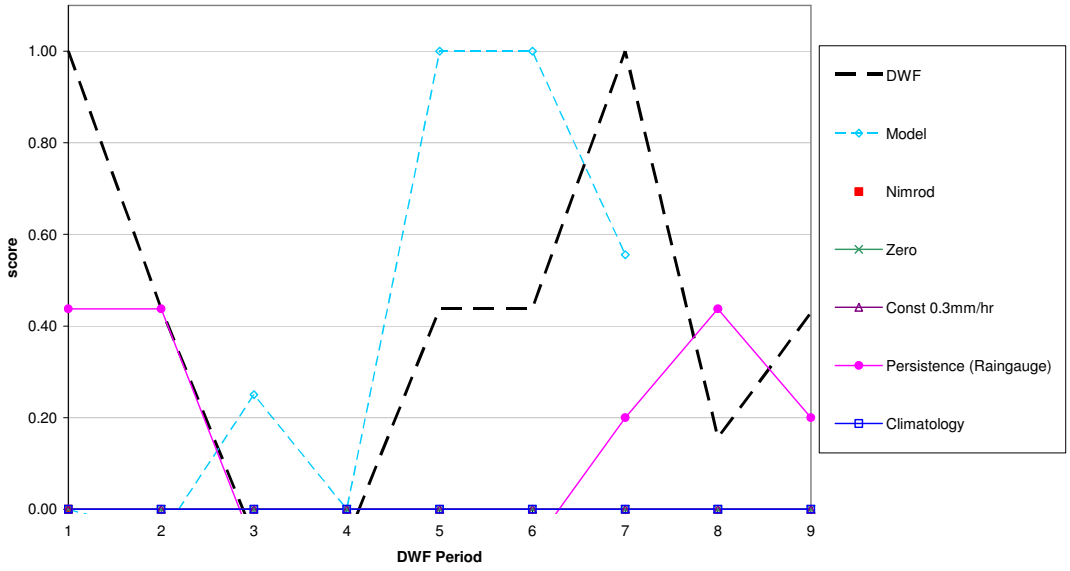
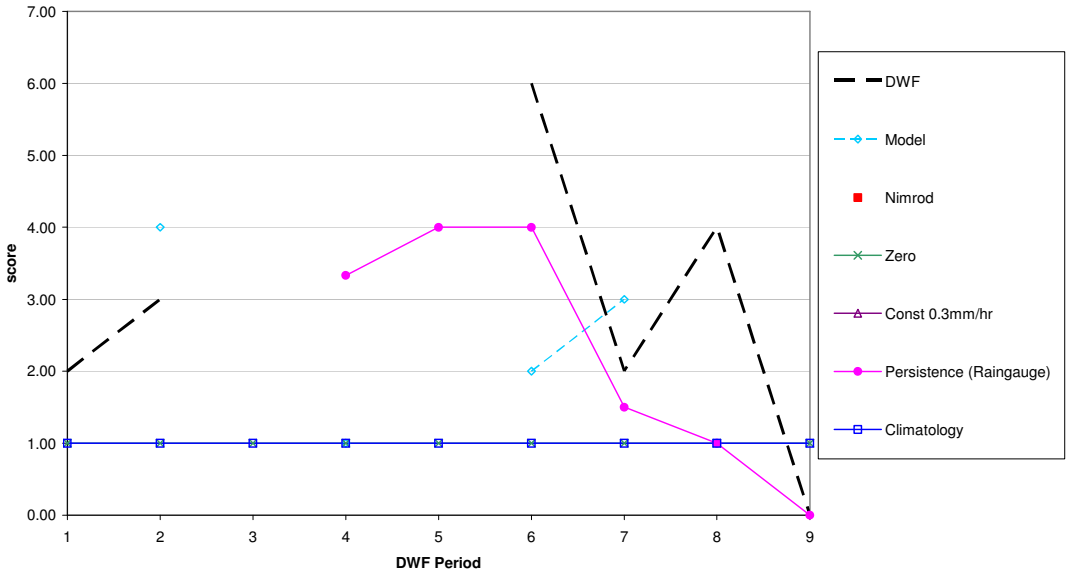


Figure 5.3.3.6 cont' Skill scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" area. Case study with 9 assessment occasions.

(i) LR1 threshold = 0 mm



(j) LR1 threshold = 4 mm

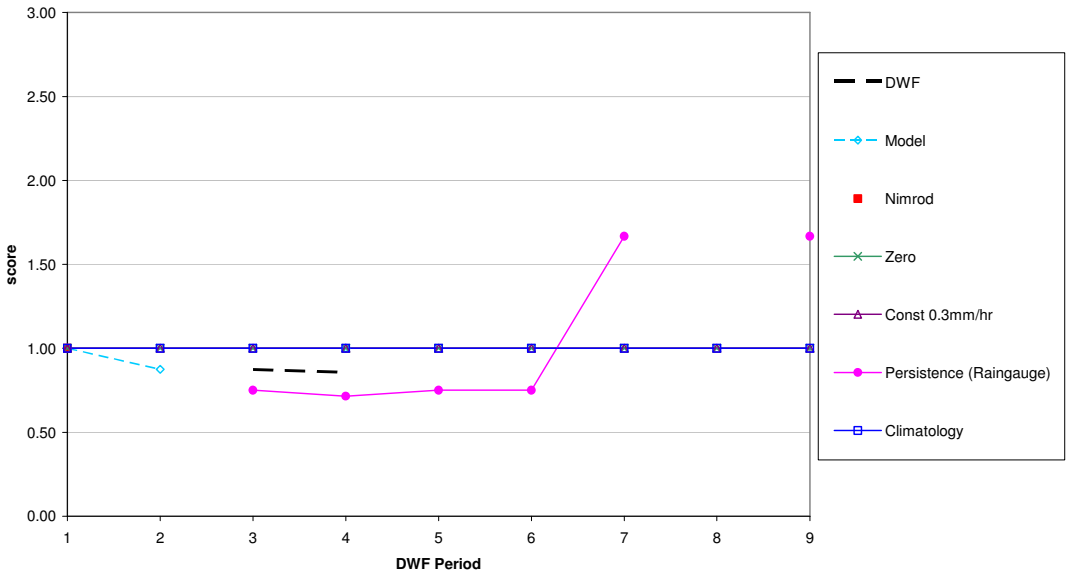
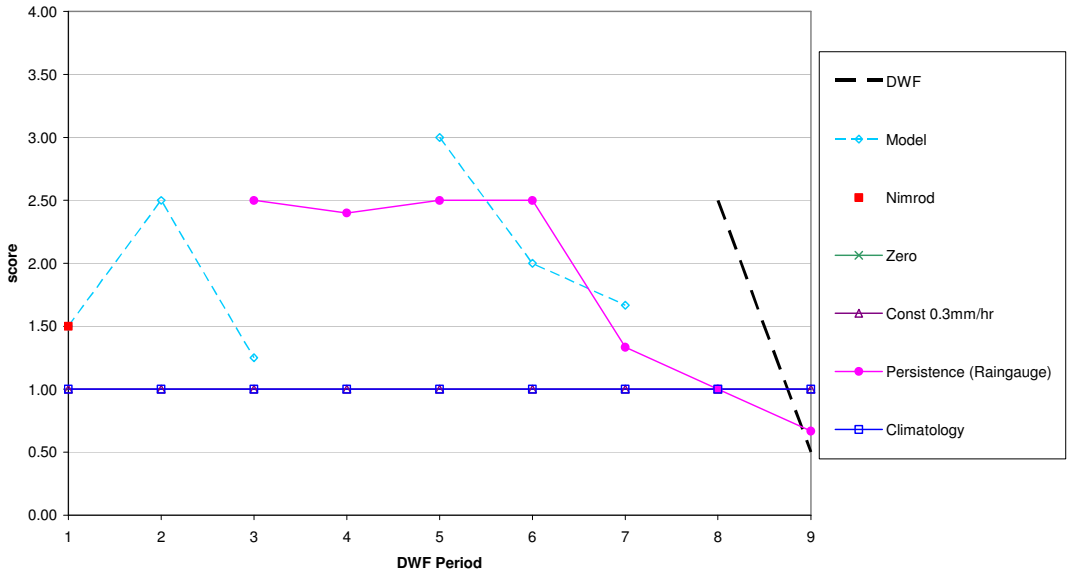


Figure 5.3.3.6 cont' Skill scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" area. Case study with 9 assessment occasions.

(k) LR2 threshold = 0 mm



(l) LR2 threshold = 4 mm

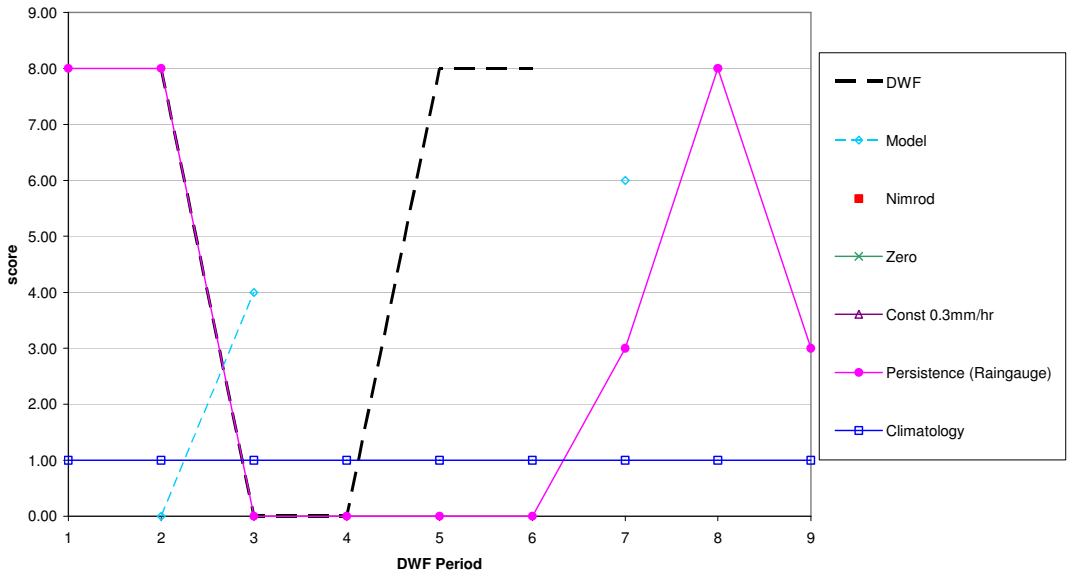
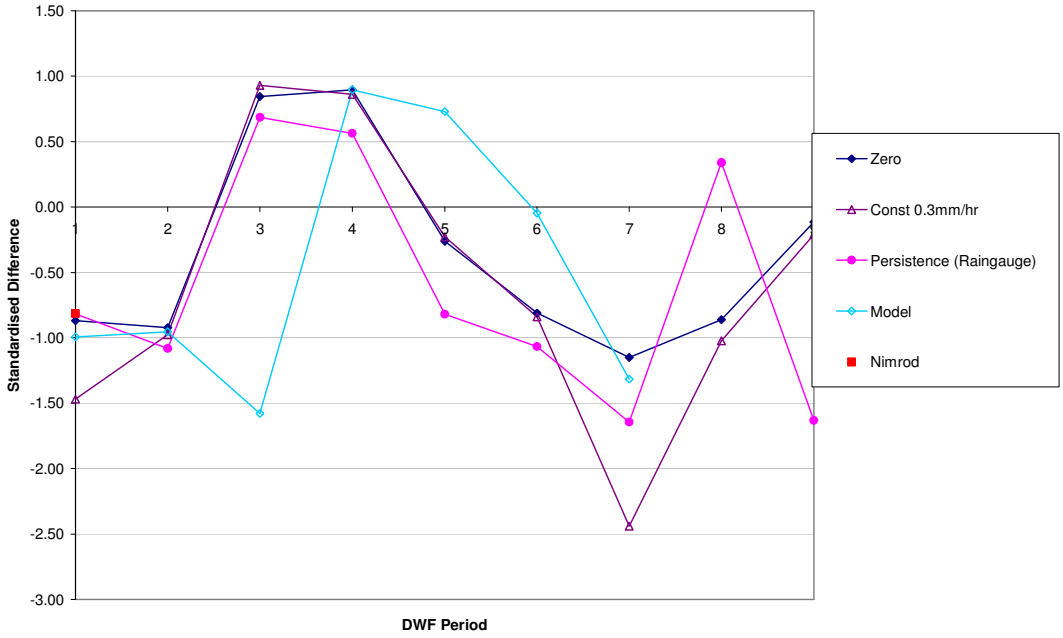


Figure 5.3.3.6 cont' Skill scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. North East Region "North East Coast" area. Case study with 9 assessment occasions.

5.3.3.7 Comparison of Forecasts

Figure 5.3.3.7 presents the standardised differences of root mean square error and root mean square error of log rainfall for the two areas.

(a) Standardised difference of root mean square error, North East Coast Area. Positive values indicate forecast better than Daily Weather Forecast.



(b) Standardised difference of root mean square error, South Pennines Area. Positive values indicate forecast better than Daily Weather Forecast.

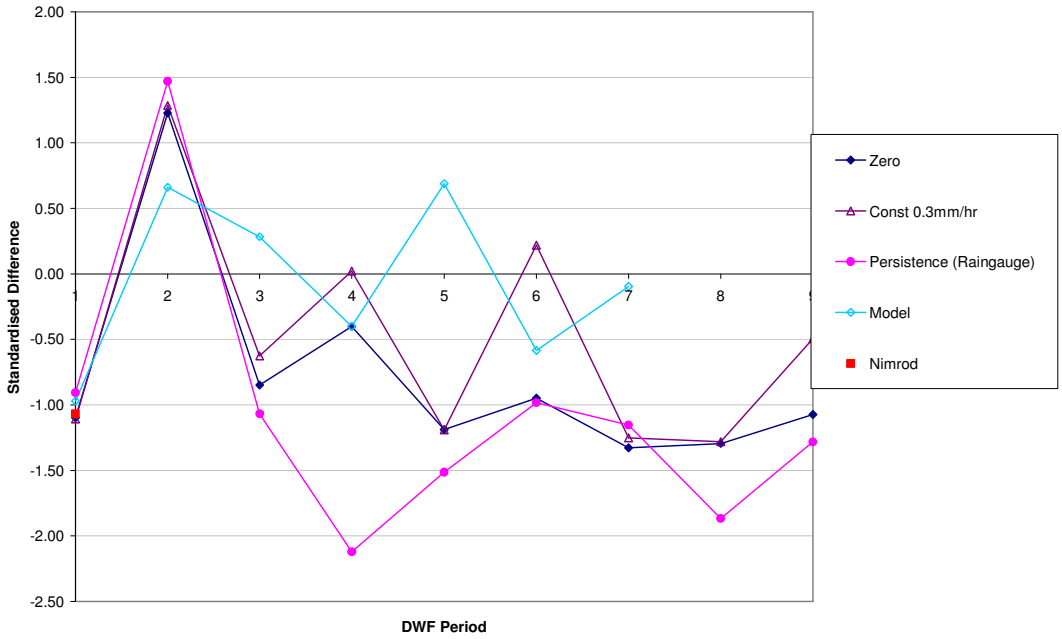
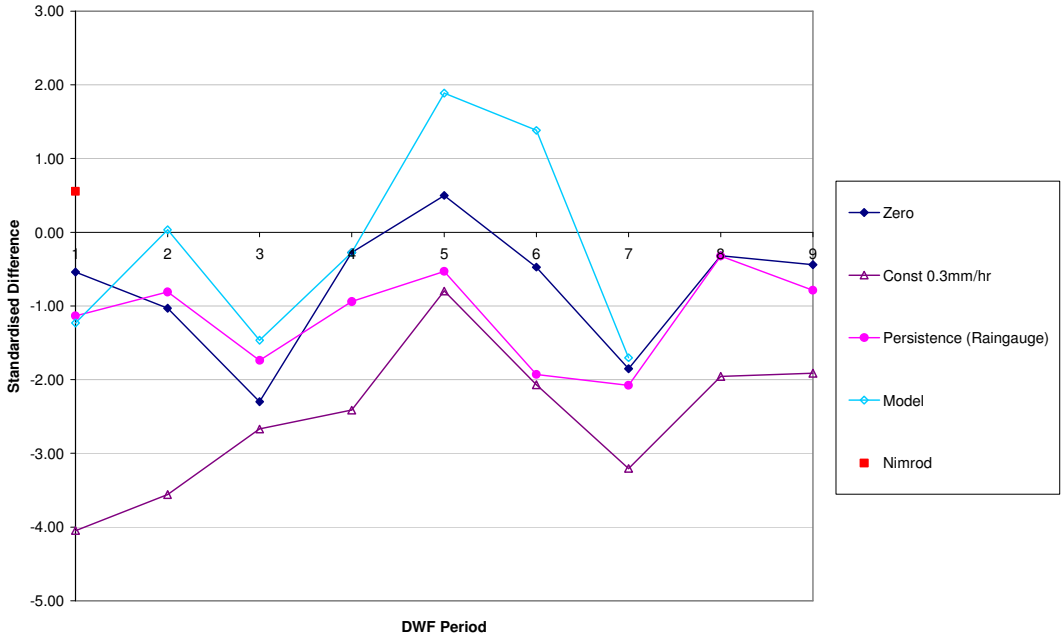


Figure 5.3.3.7 Standardised Differences of root mean square error and root mean square error of log rainfall for Daily Weather Forecast against comparative forecasts. Results shown for North East Region North East Coast and South Pennines areas, obtained using mean raingauge ground truth for case study (9 forecast occasions).

(c) Standardised difference of root mean square error of log rainfall, North East Coast Area. Positive values indicate forecast better than Daily Weather Forecast



(d) Standardised difference of root mean square error of log rainfall, South Pennines Area. Positive values indicate forecast better than Daily Weather Forecast.

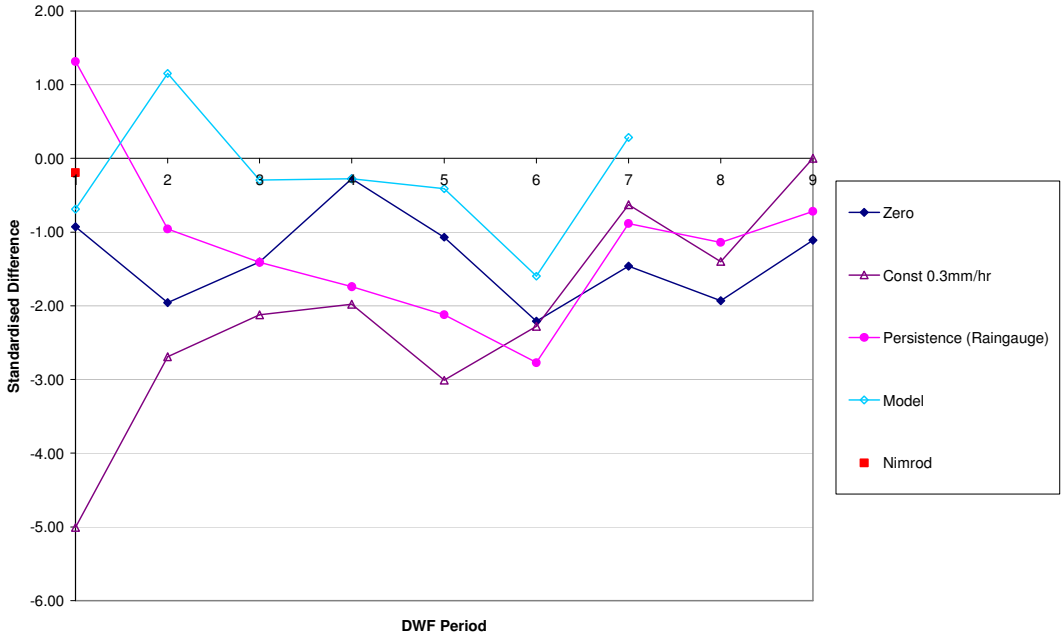


Figure 5.3.3.7 cont' Standardised Differences of root mean square error and root mean square error of log rainfall for Daily Weather Forecast against comparative forecasts. Results shown for North East Region North East Coast and South Pennines areas, obtained using mean raingauge ground truth for case study (9 forecast occasions).

5.3.4 Case Study Assessment for North West Region

5.3.4.1 Daily Weather Forecast Quantities

An example of a North West Region Daily Weather Forecast showing the relevant quantitative forecast content is given in Figure 5.3.4.1 A single section of the forecast entitled "Area Forecasts: Rainfall accumulations in mm. Days 1,2 and 3" contains the quantitative rainfall forecasts. Forecasts are given for each of the three areas for six 12 hour periods over 3 days. A Met Office document giving instructions for the construction of the Daily Weather Forecast indicates that the quantities forecast should be the spatial average accumulation over the area.

Area Forecasts: Rainfall accumulations in mm. Days 1, 2 and 3						
	Cumbria + Pennines north of the Ribble		Remainder of Lancashire		Greater Manchester, Cheshire, Merseyside	
	0000-1200	1200-2400	0000-1200	1200-2400	0000-1200	1200-2400
Tuesday	00	02	0'	05	02	09
Wednesday	03	05	10	07	14	04
Thursday	01	02	00	01	01	01

Figure 5.3.4.1 Section of Daily Weather Forecast for Northwest Region containing quantitative rainfall forecasts.

Table 5.3.4.1 summarises the forms of ground truths and comparative forecasts considered in this assessment. Due to the Daily Weather Forecast period length of 12 hours for Northwest Region, Nimrod forecast accumulations which only extend out to 6 hours could not be used as a comparative forecast source. Although the case study assessment was carried out for all three Daily Weather Forecast areas in the region, only the results for the "Cumbria and Pennines north of the Ribble" area are presented here. This area has been selected since Cumbria is mentioned in Table 5.2.1 as suffering flooding during the event used in this case study assessment.

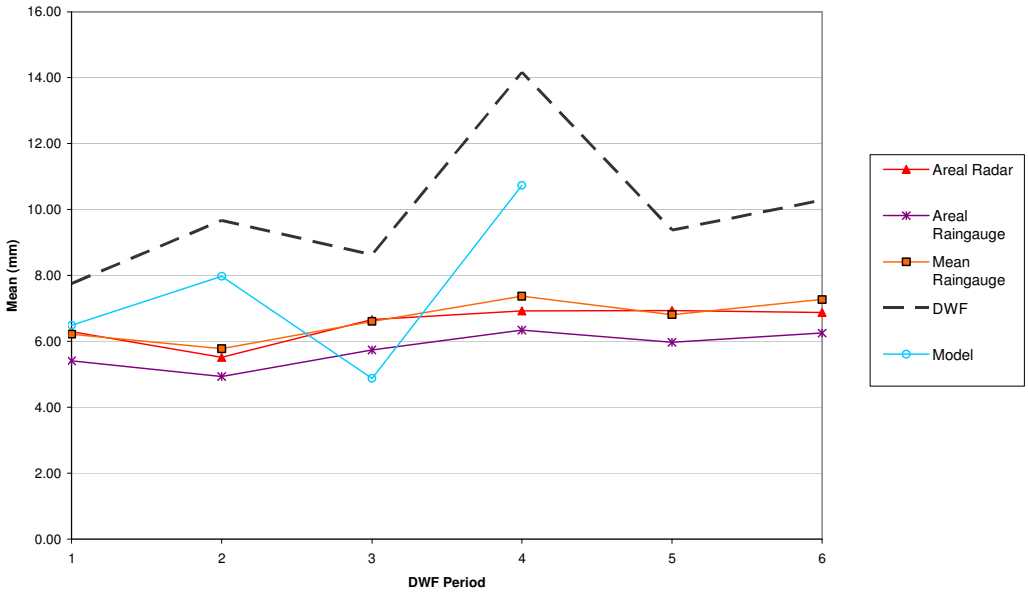
Table 5.3.4.1 Summary of target quantities, ground truths and comparative forecasts for Northwest Region Daily Weather Forecasts. Averages refer to spatial averaging carried out on raingauge and radar values which have first been accumulated over the appropriate period.

Quantity	Ground truths	Comparative forecasts
Rainfall Accumulation (mm)	**Raingauge** • Mean • Multiquadric interpolated areal average **Radar** • Areal average	**Alternative forecast sources** • Mesoscale model areal average. (Days 1 and 2 only) **Naive forecasts** • Persistence based on previous 6 hours mean raingauge accumulation. • Fixed value of 0 mm. • Fixed value of 0.3mmhr^{-1} over the forecast period.

5.3.4.2 Basic Statistics of Case Study Data

Figures 5.3.4.2 (a) to (c) present the mean, median and standard deviation of the ground truth and forecast quantities used in the case study assessment.

(a) Mean of ground truths and forecast quantities across 8 assessment occasions



(b) Median of ground truths and forecast quantities across 8 assessment occasions

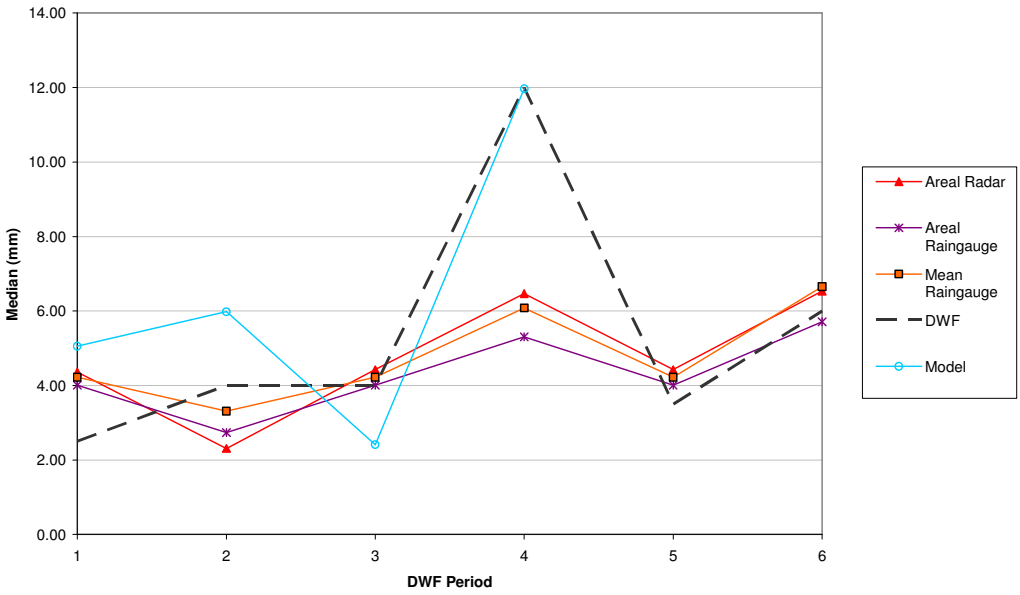


Figure 5.3.4.2 Statistics of case study forecasts and ground truths for Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

(c) Standard deviation of ground truths and forecast quantities across 8 assessment occasions

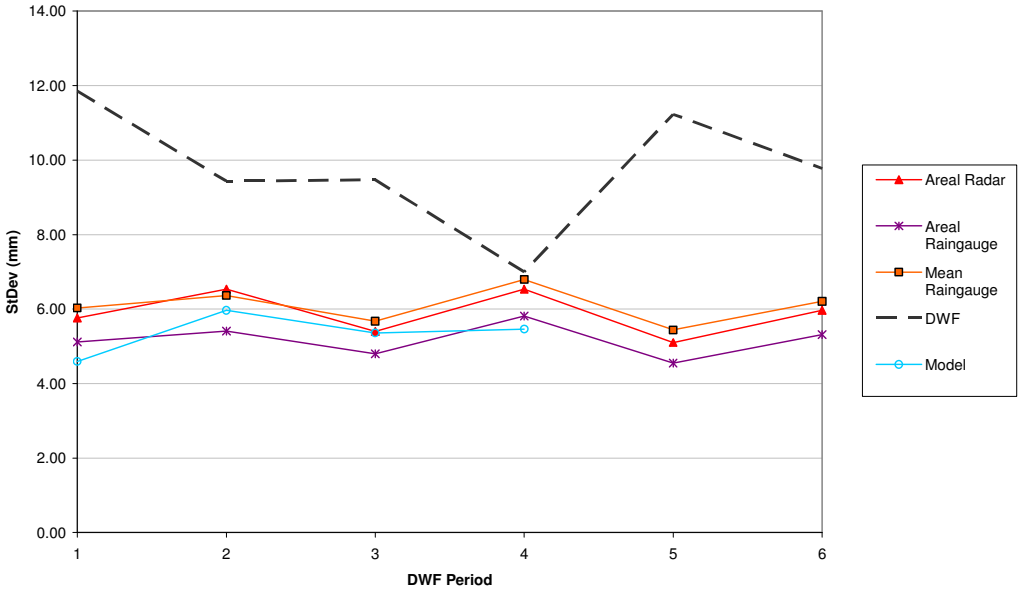


Figure 5.3.4.2 cont' Statistics of case study forecasts and ground truths for Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

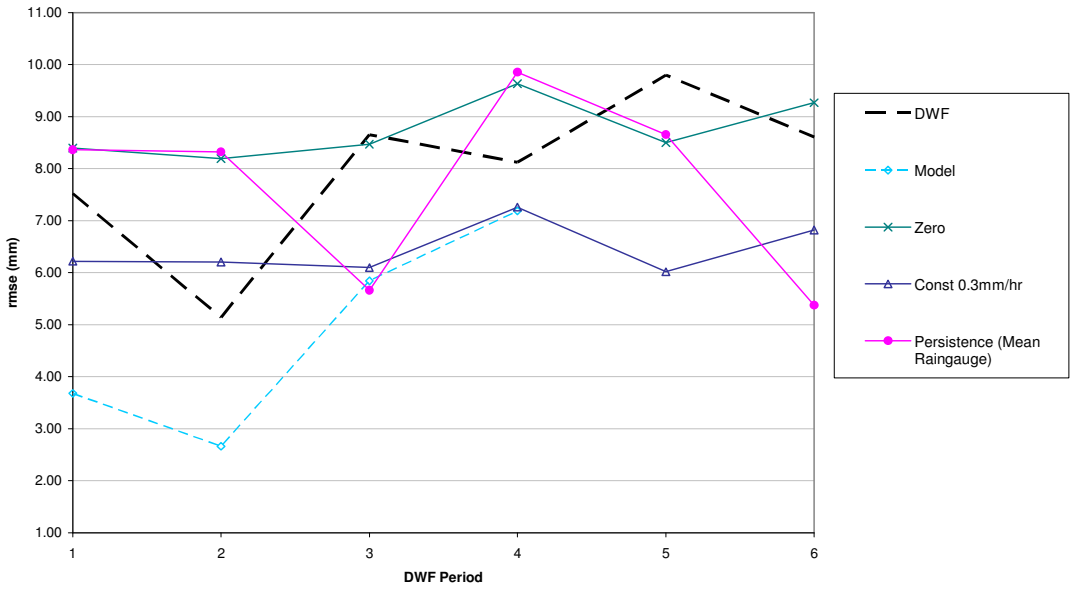
5.3.4.3 Selection of suitable forms of ground truth

The basic statistics presented in Section 5.3.4.2 indicate that, as was found for Northeast Region, the three forms of ground truth have similar statistical characteristics, at least over the case study event considered here. It therefore seems reasonable to proceed with the assessment using the mean raingauge as the ground truth.

5.3.4.4 Raw assessment measures

Figure 5.3.4.4 (a) and (b) show the root mean square error and root mean square error of log rainfall for each forecast using the mean raingauge ground truth. Both measures imply that the performance of the Mesoscale Model was at least as good or better than the Daily Weather Forecast, and that in some cases a forecast of a constant 0.3 $mmhr^{-1}$ was better than the Daily Weather Forecast. Comparison of forecasts is discussed further in Section 5.3.4.7.

(a) Root mean square error



(b) Root mean square error of log rainfall

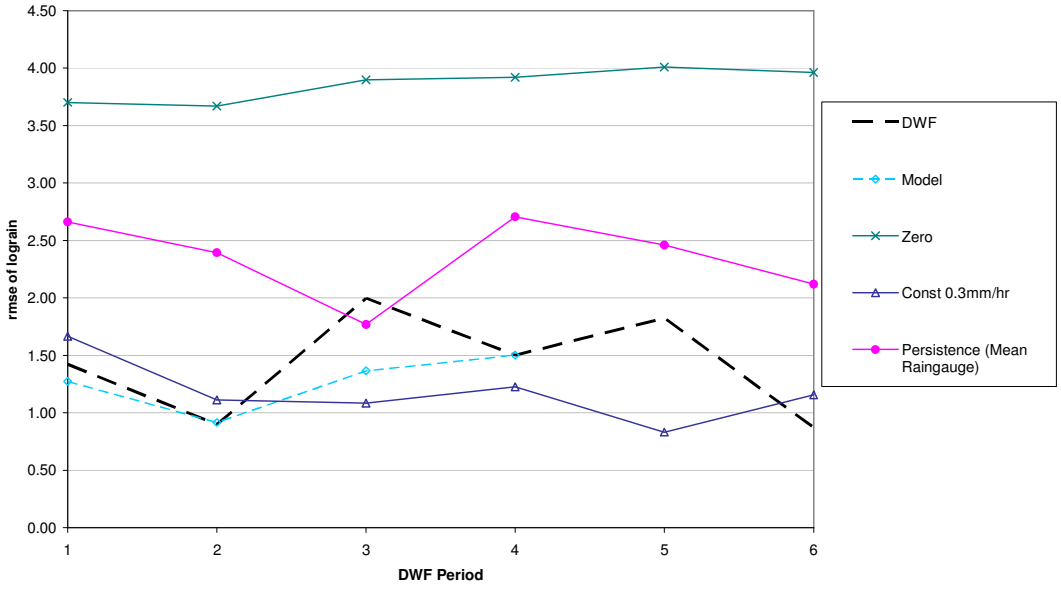
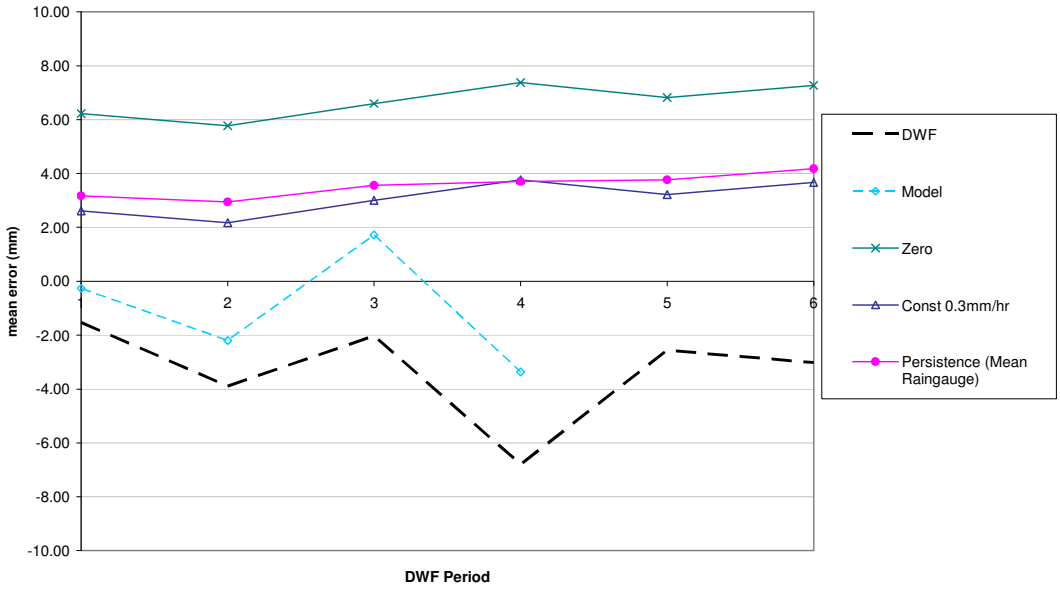


Figure 5.3.4.4 Raw performance measures for Daily Weather Forecast and comparative forecasts using mean raingauge ground truths. Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

5.3.4.5 Measures of Bias

Figure 5.3.4.5 presents bias measures for the Daily Weather forecast and comparative forecasts using the mean raingauge ground truth.

(a) Mean error



(b) Mean error of log rainfall

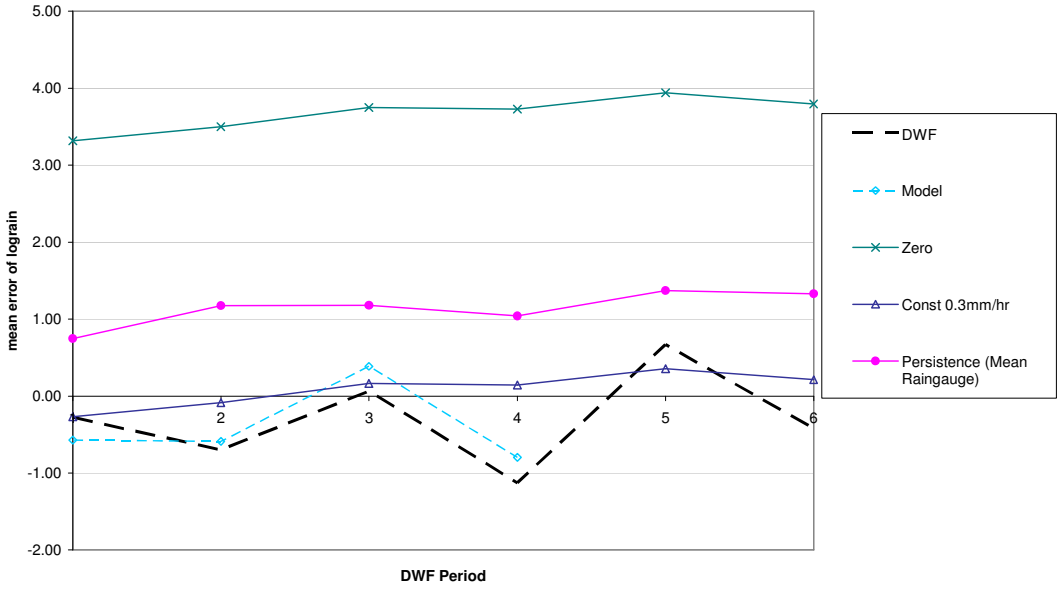
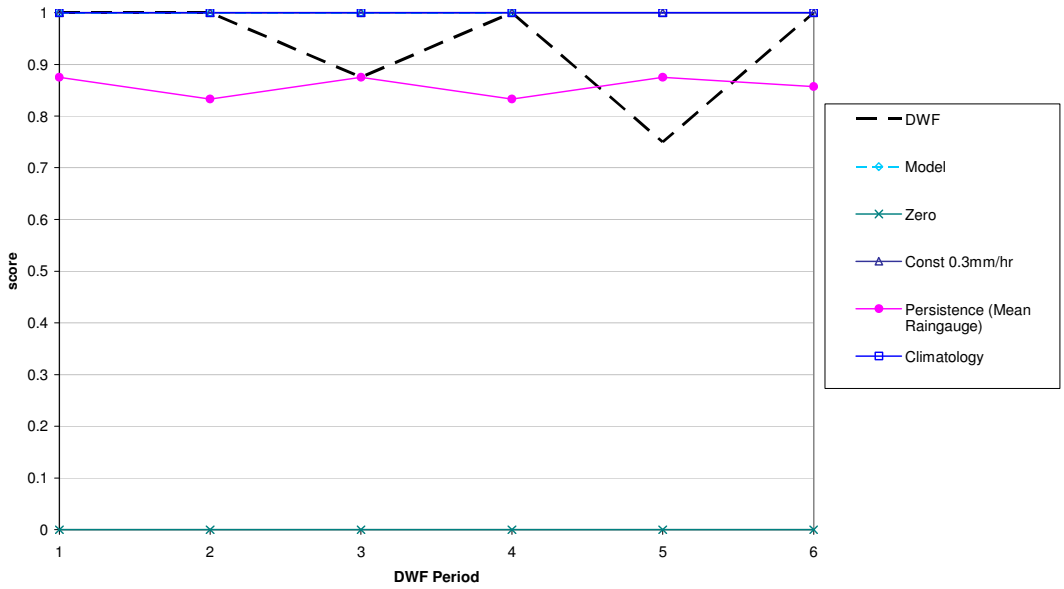


Figure 5.3.4.5 Bias measures for Daily Weather Forecast and comparative forecasts using mean raingauge ground truth. Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

5.3.4.6 Skill Scores

Figure 5.3.4.6 shows six skill scores for the Daily Weather Forecasts and comparative forecasts using the mean raingauge ground truth.

(a) CSI for threshold = 0 mm



(b) CSI for threshold = 4mm

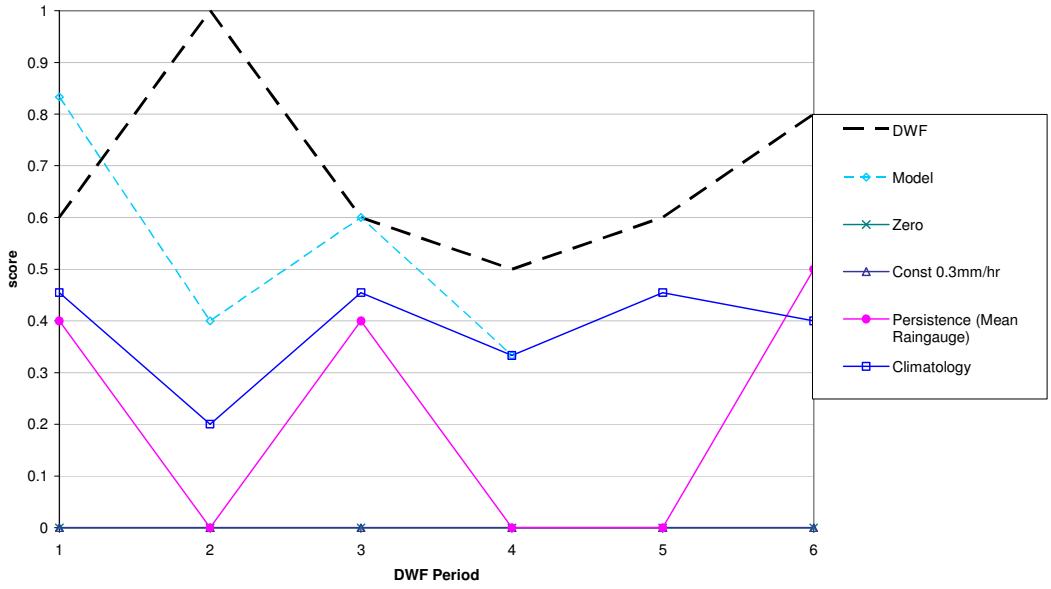
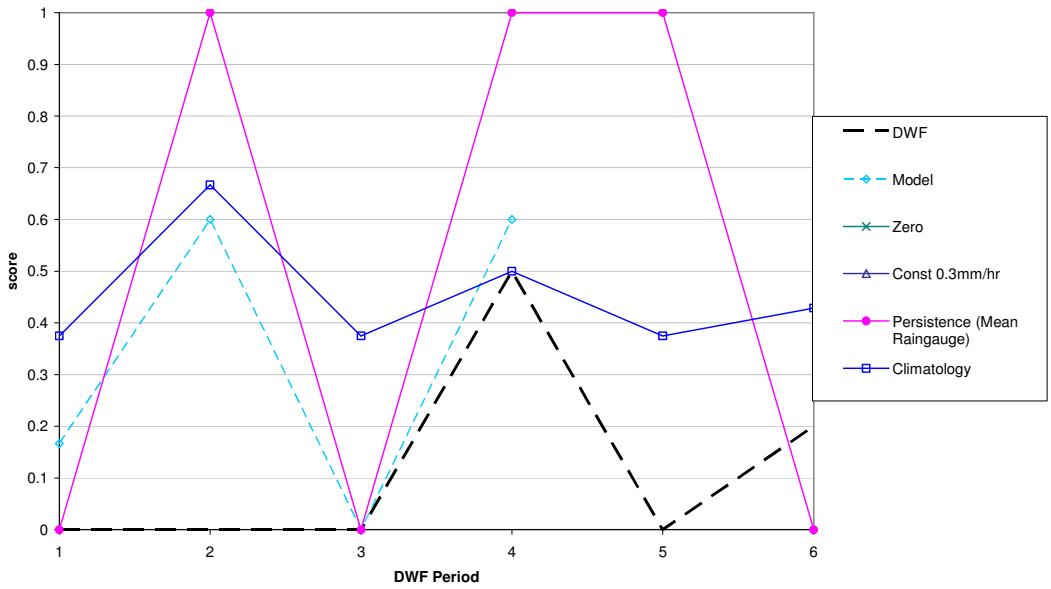


Figure 5.3.4.6 Skill Scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truths. Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

(c) FAR for threshold = 4mm



(d) POD for threshold = 4mm

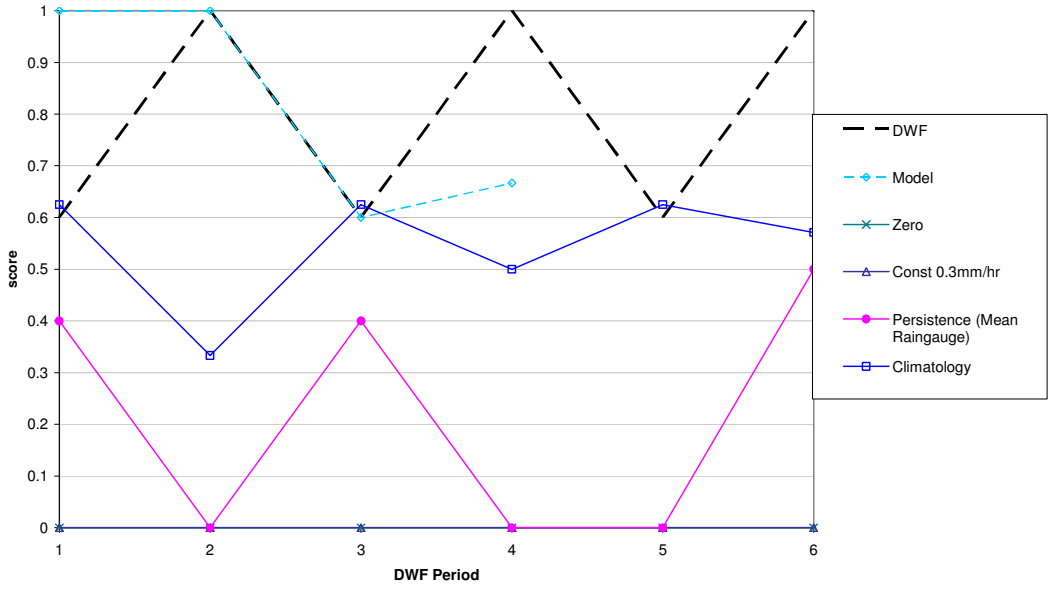
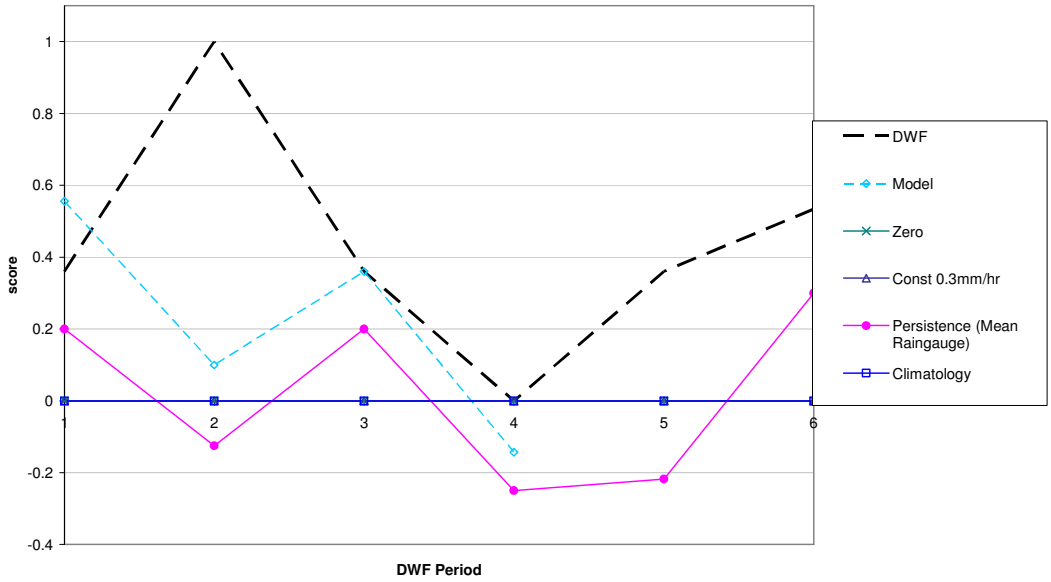


Figure 5.3.4.6 cont' Skill Scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truths. Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

(e) ETS for threshold = 4mm



(f) LR1 for threshold = 4 mm

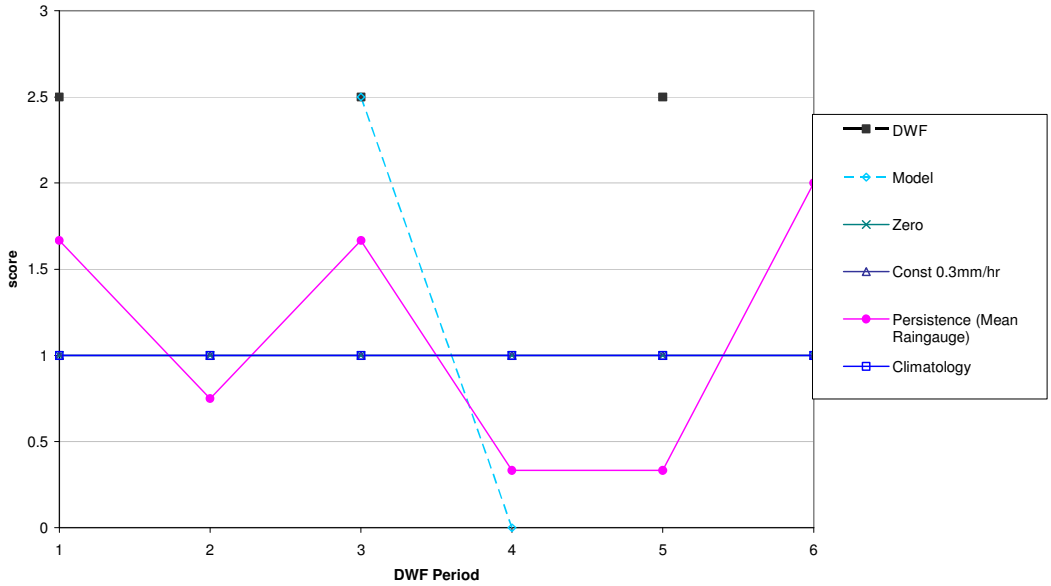
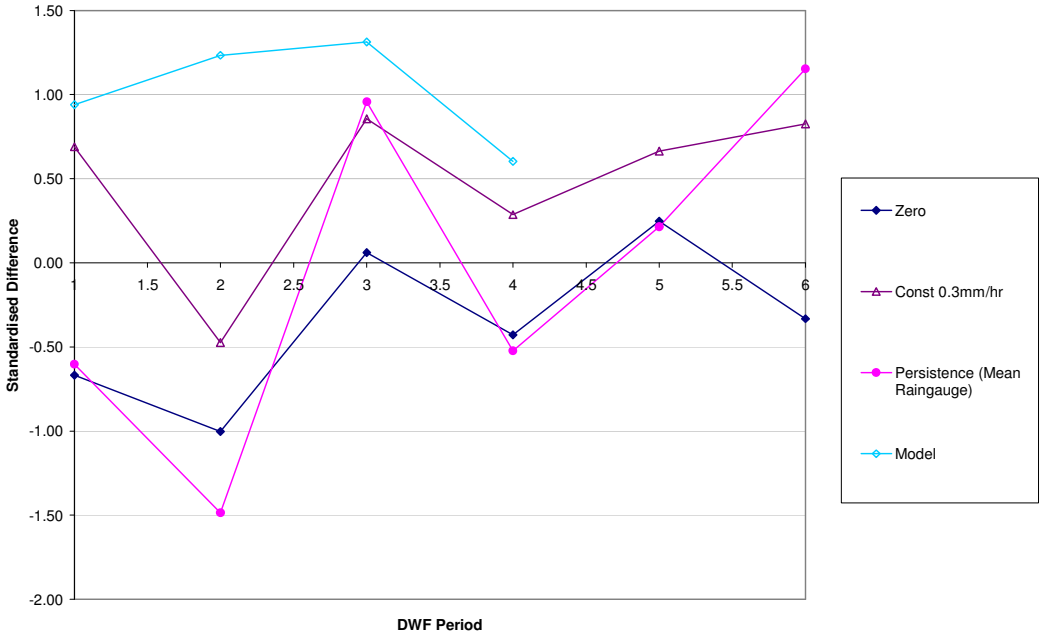


Figure 5.3.4.6 cont' Skill Scores for Daily Weather Forecast and comparative forecasts using mean raingauge ground truths. Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

5.3.4.7 Comparison of Forecasts

Figure 5.3.4.7 shows standardised differences of root mean square error and root mean square error of log rainfall for the comparative forecasts against the Daily Weather Forecast.

(a) Standardised difference of root mean square error. Positive values indicate forecast better than Daily Weather Forecast.



(b) Standardised difference of root mean square error of log rainfall. Positive values indicate forecast better than Daily Weather Forecast.

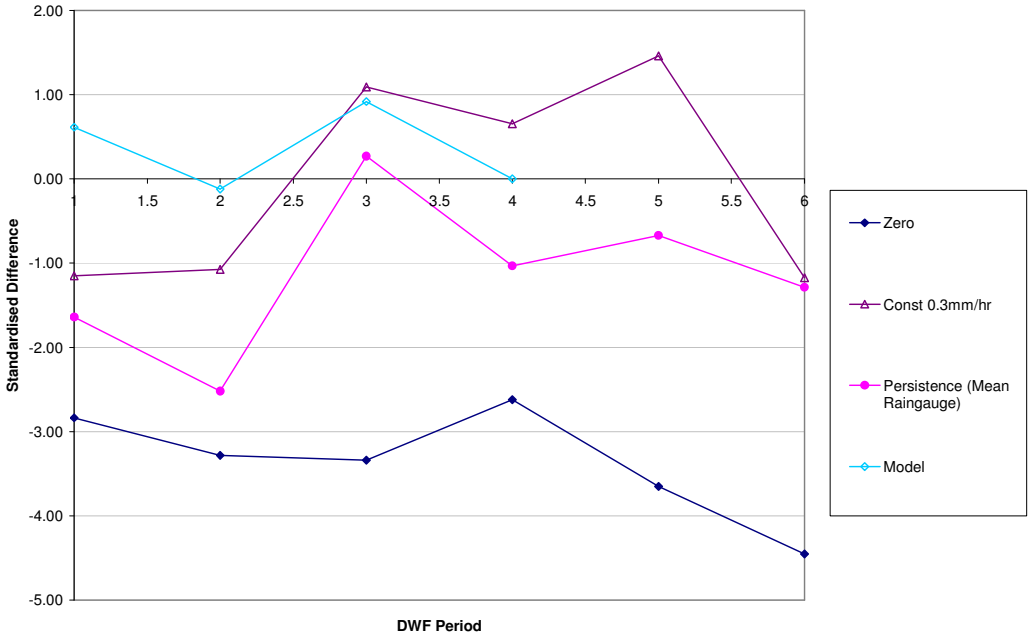


Figure 5.3.4.7 Standardised differences of raw performance measures, showing performance of forecasts compared to Daily Weather Forecast using mean raingauge ground truths. Northwest Region Area 1: Cumbria and Pennines north of the Ribble.

5.3.5 Summary

Case study assessment of the Daily Weather forecasts for Thames, Northeast and Northwest regions has been carried out using a variety of performance measures. In each case the largest number of ground truths and comparative forecasts were assessed for single case study events, consisting of 12, 9 and 8 days respectively for the three regions.

Results for Thames Region indicated that the larger number of ground truth quantities assessed could be reduced to a smaller representative set, but there was no one obvious interpretation of the "Typical" rainfall quantity. The "Max" rainfall quantity was shown to be possibly overestimated by the radar ground truth. For the other regions the different forms of ground truth gave similar results and so the mean raingauge truth was used for simplicity.

Computation of a set of raw performance measures highlighted the difference in ranking of forecasts obtained using the normal and log versions of the performance measures for the Thames Region assessment, although the differences were not as noticeable for the other regions. A number of bias measures and Skill Scores were also illustrated. Comparison of forecasts was carried out using the Standardised Difference method, which proved to be a useful way to determine the evidence for a better performance of one forecast over another. The results showed that for the small number of forecast occasions used here there was often a lack of strong evidence to prefer one forecast over another.

5.4 Assessment of Evening Updates

5.4.1 Approach to Assessment

The forecasts provided for the Evening Update service can be characterised as follows. The Evening Updates are issued on a regular basis at about 4pm each day, and cover a single fixed 18-hour time-period from 18:00 to 12:00 on the following day. Forecasts are provided for two target quantities: the largest 18 hour rainfall accumulation within an area and the highest rainfall intensity within an area over the 18-hour period. These quantities are forecasted for each of 3 areas sub-dividing each of the 3 Regions that receive Evening Updates. Besides giving values for the "most likely" outcomes of the two quantities, the forecasts include brief tables expressing the probabilities that selected threshold values will be exceeded. In both instances, the forecasts relate specifically to spatial maxima rather than to spatial averages.

The availability in text-file form of the forecast information for Thames Region has led to this Region being selected for this part of the case study. Although the formats of the files have changed over the various event periods, it has proven possible to adopt an automatic procedure which, in principle, allows all of the time-periods in Table 5.2.2 to be included in an overall assessment of performance. This gives a total of 82 occasions when the forecasts provided in the Evening Updates can be compared with the eventual outcomes. However, examination of the forecasts suggested that there had been a problem in interpretation of what was required for the forecasts of rainfall intensities until July 2002, and hence the assessment of the forecasts of rainfall intensity has been based on the forecasts from this time onwards only, giving 53 occasions when a comparison can be made between the forecasts and the eventual outcomes. The problem with the forecasts for rainfall intensities was suggested by the text associated with these entries on the forecast schema being the same as that for the rainfall amounts: specifically "most likely maximum rainfall". In practice the values for the "most likely" rainfall intensity were identical to those for the rainfall accumulations, whereas the probability tables did differ. Because of the comparisons to be made in the forecast assessments it has been simplest to treat the forecasts of rainfall intensities, including the probability forecasts, as unavailable before July 2002.

Ground Truth

For this case study, the principal source of "ground truth" data has been derived from the network of telemetering raingauges used for operational flood forecasting within the Thames Region. For the three sub-areas concerned, this network provides 47, 28 and 25 raingauges in the Northeast, Southeast and West areas of the Region, respectively. The areas of these sub-divisions are 3224, 3504 and 6190 km^2. The data were provided as 15 minute accumulations and were processed to form the 18 hour accumulations and maximum rainfall intensities for each gauge, from which the spatial maxima were formed. Given this source of ground-truth data, the rainfall intensities derived relate to average intensities over 15 minute time-periods.

An additional source of ground-truth data for this case study is weather radar. Notionally, this might provide a better source of ground-truth data than the raingauge network because of its superior spatial coverage. However, quantitative estimates of

rainfall from weather radar are not always reliable. Within the time-scale of the present phase of the project we have not been able to implement an automatic procedure to derive the spatial maximum of the 18-hour total rainfalls derived from radar sources, but we have been able to derive the values for the maximum rainfall rates within an area. We are therefore able to compare the results of the forecasts of maximum rainfall intensity against both raingauge and radar sources. It seems that the best source of radar- derived rainfall information that will be available in near real-time will be the Nimrod quality controlled "actual" product. Hence it seems useful to undertake a comparison of the forecasts from the Evening Updates against this data source. The comparison made here should be treated with caution because this particular radar-product was still under operational development during the time-period used for the assessment. In particular, full sets of quality control procedures may not have been in place, and the availability of raingauge information for adjustment is unclear: either of these two aspects of the Nimrod product may have changed during the assessment period.

The precise definition of the target for the forecast of rainfall intensity is unclear, but discussions with EA staff have indicated that they interpret these values in relation to what might have been seen in a radar-based rainfall display of rainfall rates. Such display values are based on instantaneous snapshots of rainfall intensity made at either a 5 or 15 minute time-interval. The Nimrod rainfall product is available at a 15-minute time-step, in a form which is a composite of 1, 2 and 5km resolutions. The quantity derived from the Nimrod product for comparison against the rainfall intensity component of the Evening Update was the maximum of all the 15-minute rainfall values falling within the 18-hour forecast period and within the particular sub-area of the Thames Region (for 1 km pixels entirely within the sub-area).

Comparative Forecasts

Nominally the Evening Updates provide two separate forecasts for both maximum rainfall accumulation and maximum rainfall rate, one of which is an ordinary, single-valued forecast and the other a probability forecast. However the probability forecast could potentially be used to derive other single-valued forecasts related to the location of the probability distribution. It seems from the examples available that there is no clear relationship between the "most likely value" given in the forecast and the probability forecast, but this lack of relationship may be as much due to the poor resolution of the probability forecasts as to any other underlying problem. In these circumstances it seemed reasonable to extract a single-valued forecast from the Evening Updates' probability forecast for use as a comparator forecast within the assessment procedures, and the simplest such forecast to extract was deemed to be the median of the probability distribution. (The median of a probability distribution for rainfalls is the value of rainfall such that there is a 50% or more chance that the outcome will be equal to or higher than the value, and a 50% or more chance that the outcome will be equal to or lower than the value). The precise value calculated for the median (or any other statistic) in this case depends upon the interpretation placed on the tables of probabilities when they are used to define an overall probability distribution function. The method used was based on linear interpolation in the tables, rather than fitting some parametric distribution.

The following is an example of the calculation of the median value from the probability forecast. The, the probability forecast in the Evening Update for the Northeast area of Thames Region on 29 July 2002 (for the maximum rainfall amount) was as follows.

Probability of Amount > 0 mm :	60%
Probability of Amount > 10 mm :	20%
Probability of Amount > 20 mm :	10%
Probability of Amount > 40 mm :	0%

Most likely amount : 12 mm.

The median is the value which has a 50% chance of being exceeded, and according to the above table, this value must be between 0 and 10. The 50% point is one quarter of the way from 60% to 20%, and hence the estimated value for the median is one quarter of the way from 0 mm to 10mm. Thus the median is calculated as 2.5 mm.

The networked radar products do not include forecasts of rainfall out to a lead-time of 18-hours and hence they do not provide a source of comparative forecasts. While the mesoscale model may eventually provide a possible alternative source of forecasts, data from this source were not available for the present phase of the study.

To summarise, for this phase of the project, the main set of forecasts that are available for comparison are all derived from the Evening Updates and are:

(i) the explicit forecast indicated as “most likely value”;
(ii) a derived forecast, calculated as the median of the probability forecast;
(iii) the probability forecast itself.

Thus there are two single-valued forecasts and one probability forecast. It seemed reasonable to extend this set of candidates in two ways. Firstly, by defining some additional single-valued forecasts of a rather simple nature and, secondly, by defining some additional probability forecasts which can be derived from the single-valued forecasts in a simple way. It is convenient to treat the assessment of single-valued forecasts and probability forecasts as separate tasks, but it should be noted that among the simple probability forecasts are some which correspond to expressing absolute certainty about a single value.

The simple single-valued forecasts that have been included for comparison are of two types. For the first type, the forecast is based on recently observed values of the same type as that which are to be forecasted. Given that the time-period of the forecasts here are 18-hours, it is convenient to construct forecasts from observed values over a corresponding 18-hour time-period. One forecast is constructed from values observed in the 18-hour period immediately before the start of the forecast period. A second forecast is constructed from values observed in the 18-hour period starting 24 hours before the start of the period. The second type of single-valued forecast is constructed by using a single constant value for the forecast: the assessment has been performed for two different constant values for each of the quantities being forecasted.

The simple probability forecasts included for comparison are of two types. For the first type, the single-valued forecasts outlined above are included with the probability component of the forecast constructed so as to express absolute certainty in the single-value forecast. The second type of probability forecast is again constructed from the single-valued forecast, but with the uncertainty in the forecast being determined by the rule that the probability is uniformly distributed over an interval centred on the

single-valued forecast with a width that is the same as the central value (i.e. from 0 to 200% of the central value), with an overriding minimum of 1 unit (mm or mm h^{-1}, depending on the quantity being forecasted). In instances where this interval extends to negative values, the probability distribution is revised so that the probability for negative values is replaced by a discrete component of probability for the zero value. The choice of the size of the interval used here is entirely arbitrary and there might be better ways of associating a probability with the single-valued forecasts.

Table 5.4.1.1 provides a summary of the ground-truth and comparative forecasts that are available for this study for the 18-hour accumulation component of the Evening Update forecast. Table 5.4.1.2 provides a similar summary for the maximum rainfall rate component of the Evening Updates.

Table 5.4.1.1 Summary of Assessment for Evening Update forecasts of Maximum Rainfall Accumulations

Description	Abbreviation
Ground truth	
Maximum 18-hour accumulation across raingauges in area	
Single-valued forecasts	
Operational candidates	
Values labelled 'most likely' in Evening Update	Most Likely
Median of probability forecast in Evening Update	Prob. Median
Comparative forecasts	
Maximum 18-hour raingauge accumulation for period starting 18 hours before initial forecast time	$Persist_{RG,18}$
Maximum 18-hour raingauge accumulation for period starting 24 hours before initial forecast time	$Persist_{RG,24}$
A fixed value of zero mm for the maximum accumulation	$Const_{0mm}$
A fixed value of 5 mm for the maximum accumulation	$Const_{5mm}$
Probability forecasts	
Operational candidates	
Probability Forecast from Evening Update	Prob. Forecast
Comparative forecasts	
The single-valued forecasts listed above treated as being absolutely certain	(certain)
The single-valued forecasts listed above, with uncertainty uniform over ± 100% or ± 1mm, whichever is larger.	(100% error)

Table 5.4.1.2 Summary of Assessment for Evening Update forecasts of Maximum Rainfall Rates

Description	Abbreviation
Ground truth	
Maximum of all 15-minute accumulations at raingauges in the area in the 18-hour period, converted to rate	
Maximum 15-minute rainfall rate in the 18-hour period in the area as estimated by the Nimrod radar product	
Single-valued forecasts	
Operational candidates	
'Most likely' from Evening Update	Most Likely
Median of probability forecast in Evening Update	Prob. Median
Comparative forecasts	
Maximum of all 15-minute accumulations at raingauges in area in the 18-hour period starting 18 hours before initial forecast time, converted to rate	$Persist_{RG,18}$
Maximum of all 15-minute accumulations at raingauges in area in the 18-hour period starting 24 hours before initial forecast time, converted to rate	$Persist_{RG,24}$
Maximum of all 15-minute rainfall rates, as estimated by the Nimrod radar product, in the area in the 18-hour period starting 18 hours before initial forecast time	$Persist_{RD,18}$
Maximum of all 15-minute rainfall rates, as estimated by the Nimrod radar product, in the area in the 18-hour period starting 24 hours before initial forecast time	$Persist_{RD,24}$
A fixed value of zero mm h^{-1} for the maximum rate	$Const_{0mm/hr}$
A fixed value of 10 mm h^{-1} for the maximum rate	$Const_{10mm/hr}$
Probability forecasts	
Operational candidates	
Probability Forecast from Evening Update	Prob. Forecast
Comparative forecasts	
The single-valued forecasts listed above treated as being absolutely certain	(certain)
The single-valued forecasts listed under (i) to (viii) with uncertainty uniform over ± 100% or ± 1mm h^{-1}, whichever is larger.	(100% error)

5.4.2 Example Forecasts and Outcomes

Table 5.4.2.1 lists the full set of data for the assessment of forecasts for the North East area of the Agency's Thames Region in the case of the maximum 18-hour accumulation forecast. The dates and times reported here indicate the start of the forecast period. Times have been converted to GMT.

Table 5.4.2.1 Example of data for assessment of rainfall forecasts for 18-hour rainfall accumulations: maximum totals in Northeast area of Thames Region (units: mm).

date	--- Evening Update ---		--- Persistence ---		Outcome from
	'most likely'	Median	18 hour	24 hour	Raingauges
21 1 2002 18:00	4.00	5.00	0.80	2.20	2.00
22 1 2002 18:00	10.00	8.33	2.00	2.00	6.40
23 1 2002 18:00	10.00	10.00	8.00	6.40	4.80
24 1 2002 18:00	4.00	6.00	3.00	4.80	0.80
25 1 2002 18:00	6.00	5.56	4.00	0.80	19.20
27 1 2002 18:00	4.00	7.14	3.80	6.40	5.00
28 1 2002 18:00	0.50	0.00	5.00	5.00	0.00
29 1 2002 18:00	1.00	4.44	0.20	0.00	0.20
30 1 2002 18:00	8.00	6.67	3.80	0.20	4.60
31 1 2002 18:00	8.00	7.14	3.00	4.60	6.00
1 2 2002 18:00	3.00	5.00	1.80	6.00	0.40
2 2 2002 18:00	3.50	3.33	0.20	0.40	1.80
3 2 2002 18:00	12.50	10.00	7.80	1.80	18.20
5 2 2002 18:00	4.00	5.56	5.60	3.80	3.80
6 2 2002 18:00	1.00	0.00	0.20	3.80	1.60
7 2 2002 18:00	2.00	5.00	3.20	1.60	1.20
8 2 2002 18:00	2.00	5.00	1.20	1.20	1.20
9 2 2002 18:00	3.00	0.00	2.20	1.20	1.80
10 2 2002 18:00	12.00	12.50	0.20	1.80	4.20
11 2 2002 18:00	1.00	2.00	6.40	4.20	3.80
9 6 2002 17:00	6.00	4.00	3.40	0.40	3.40
10 6 2002 17:00	2.00	1.72	6.80	3.40	6.80
11 6 2002 17:00	2.00	3.75	0.20	6.80	10.60
12 6 2002 17:00	2.00	3.75	6.80	10.60	3.80
13 6 2002 17:00	2.00	1.67	3.80	3.80	0.40
14 6 2002 17:00	5.00	1.67	1.00	0.40	0.20
15 6 2002 17:00	5.00	5.00	0.00	0.20	1.60
16 6 2002 17:00	0.00	0.00	1.60	1.60	0.20
17 6 2002 17:00	2.00	2.86	0.20	0.20	19.40
25 7 2002 17:00	0.00	0.00	0.20	0.20	0.00
26 7 2002 17:00	0.00	0.00	0.00	0.00	0.00
27 7 2002 17:00	0.00	0.00	0.00	0.00	0.00
28 7 2002 17:00	5.00	0.00	0.00	0.00	0.00
29 7 2002 17:00	12.00	2.50	2.00	0.00	9.60
30 7 2002 17:00	3.00	2.00	13.60	9.60	44.20
31 7 2002 17:00	10.00	10.00	33.80	44.20	5.00
1 8 2002 17:00	2.00	0.00	0.20	5.00	0.20
2 8 2002 17:00	7.00	3.33	0.20	0.20	4.60
3 8 2002 17:00	10.00	7.50	13.40	4.60	23.00
4 8 2002 17:00	15.00	12.00	22.40	23.00	1.80
5 8 2002 17:00	18.00	14.00	8.40	1.80	4.40
6 8 2002 17:00	1.00	3.75	3.40	4.40	1.80
7 8 2002 17:00	30.00	12.50	6.80	1.80	22.40
8 8 2002 17:00	10.00	10.00	13.80	22.40	15.20
9 8 2002 17:00	12.00	12.00	28.00	15.20	24.80
10 8 2002 17:00	5.00	6.00	4.60	24.80	5.20
11 8 2002 17:00	1.00	1.67	0.20	5.20	0.40
4 9 2002 17:00	0.00	0.00	0.20	0.00	0.20
5 9 2002 17:00	0.50	0.00	0.40	0.20	2.60
6 9 2002 17:00	10.00	6.25	3.20	2.60	8.80
7 9 2002 17:00	3.00	2.00	8.80	8.80	0.60
8 9 2002 17:00	12.50	10.00	2.20	0.60	11.20
9 9 2002 17:00	12.00	10.00	44.40	11.20	6.80
10 9 2002 17:00	0.00	0.00	0.20	6.80	0.00
8 10 2002 17:00	0.50	0.00	0.00	0.00	0.20
9 10 2002 17:00	0.00	0.00	0.20	0.20	0.40

date	--- Evening Update ---		--- Persistence ---		Outcome from
	'most likely'	Median	18 hour	24 hour	Raingauges
10 10 2002 17:00	0.00	0.00	0.40	0.40	0.40
11 10 2002 17:00	8.00	5.00	0.40	0.40	11.80
12 10 2002 17:00	0.00	0.00	12.20	11.80	0.20
13 10 2002 17:00	10.00	12.00	8.60	0.20	10.00
14 10 2002 17:00	8.00	8.33	4.60	10.00	8.20
15 10 2002 17:00	8.00	8.33	33.20	8.20	4.60
16 10 2002 17:00	1.00	2.00	4.00	4.60	0.20
17 10 2002 17:00	0.00	0.00	0.20	0.20	3.20
18 10 2002 17:00	0.00	0.00	3.20	3.20	0.20
19 10 2002 17:00	2.00	5.26	0.20	0.20	0.20
20 10 2002 17:00	6.00	6.67	4.20	0.20	4.60
21 10 2002 17:00	10.00	10.00	10.80	4.60	13.00
22 10 2002 17:00	6.00	4.00	19.80	13.00	3.40
23 10 2002 17:00	1.00	0.00	15.40	3.40	0.20
24 10 2002 17:00	1.00	2.86	0.20	0.20	6.00
25 10 2002 17:00	5.00	7.14	10.60	6.00	4.20
26 10 2002 17:00	15.00	12.50	4.20	4.20	5.00
27 10 2002 18:00	0.50	0.00	4.00	5.00	0.60
28 10 2002 18:00	7.00	2.86	0.20	0.60	0.00
29 10 2002 18:00	7.00	6.67	2.00	0.00	3.80
30 10 2002 18:00	8.00	7.14	9.00	3.80	1.40
31 10 2002 18:00	1.00	1.67	0.80	1.40	0.20
1 11 2002 18:00	4.00	6.25	5.80	0.20	1.00
2 11 2002 18:00	10.00	14.00	9.00	1.00	6.60
3 11 2002 18:00	1.00	0.00	3.80	6.60	1.00
4 11 2002 18:00	2.00	1.67	0.20	1.00	0.00

The values given in Table 5.4.2.1 can be used to compare the two single-valued forecasts derived from the Evening Updates: the 'most likely' value, quoted directly in the forecast, and the median of the probability forecast. These values do tend to vary together in a reasonable way, but there are often sizeable differences. The 'most likely' value and median value of a probability distribution measure different characteristics of the distribution, and hence some differences would be expected even if the values were formally derived from a fully defined distribution. Given that the forecast values are defined in a less formal way, this would lead to greater differences. A further contributory factor is thought to be the use of a relatively imprecise way of expressing the probability forecast in the form of exceedence probabilities for only a few levels of rainfall amount, which leads to inaccuracies in deriving the median.

A simple way to assess the performance of forecasts is by visual examination of scatter plots of the forecasts and outcomes. A complete set of such scatter plots for the present case study, and for the case of forecasts of rainfall amounts, is provided in Figures 5.4.2.1 to 5.4.2.3. These plots indicate that there is not a particularly good correspondence between the operational forecasts and the outcomes as derived from the raingauge network. More importantly for the purposes of the analysis here, it is not the case that the performance analyses will be completely dominated by only one or two particularly bad forecasts.

The values forecasted for the different regions on a given occasion tend to be rather similar, but there are considerable differences in the corresponding outcomes. Figures 5.4.2.1 to 5.4.2.3 all show a single isolated relatively high rainfall outcome, but these do not all relate to the same rainfall event. The highest values for the sub-areas occurred on 30 July 2002 in the Northeast and Western sub-areas, when the values were 44.2 mm and 52.8 mm respectively, and on 7 August 2002 (41.6 mm) in the Southeast sub-area.

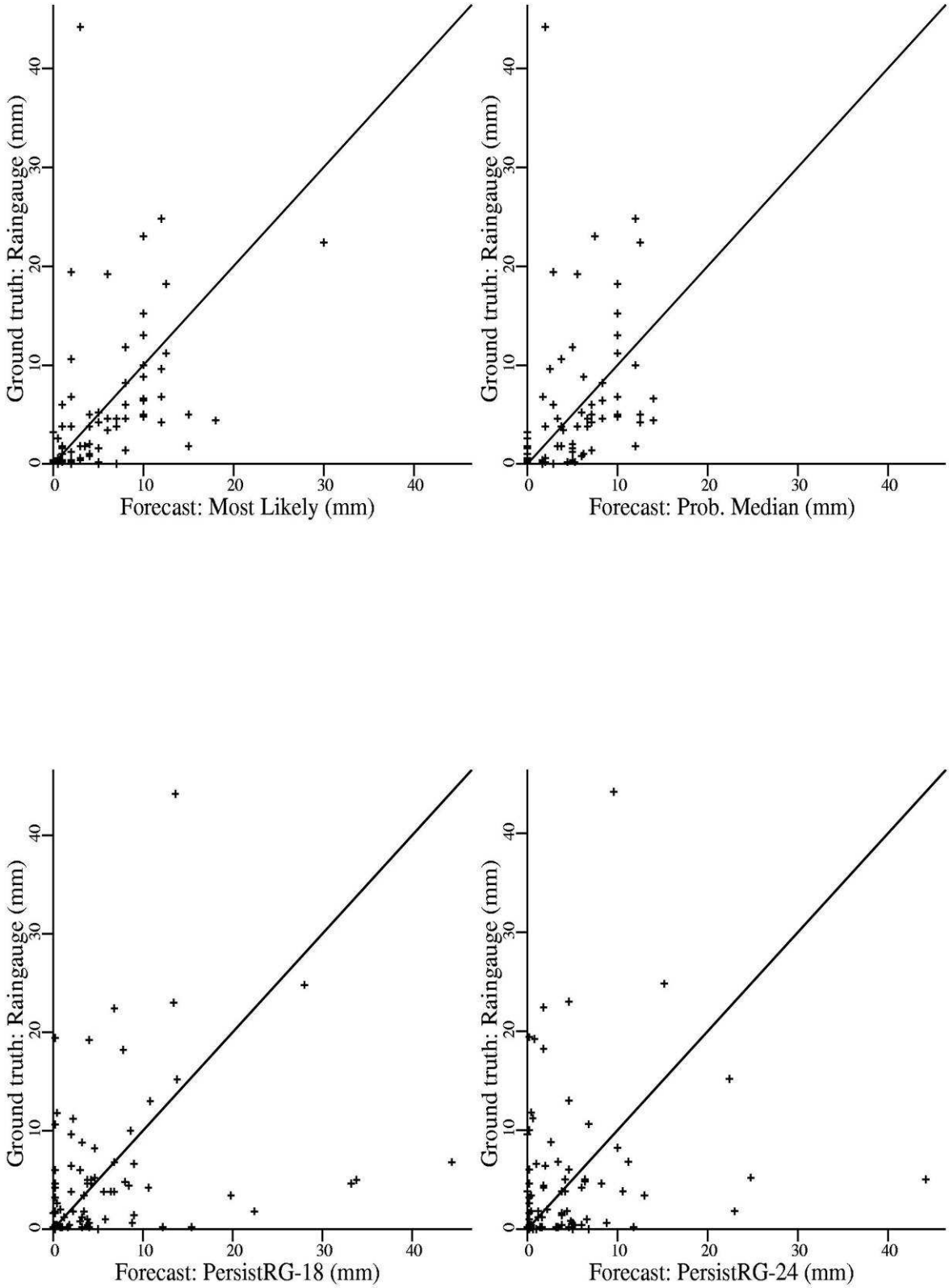


Figure 5.4.2.1 Evening Update forecasts of maximum rainfall amounts. Ground truth from raingauge network. Northeast sub-area of Thames Region

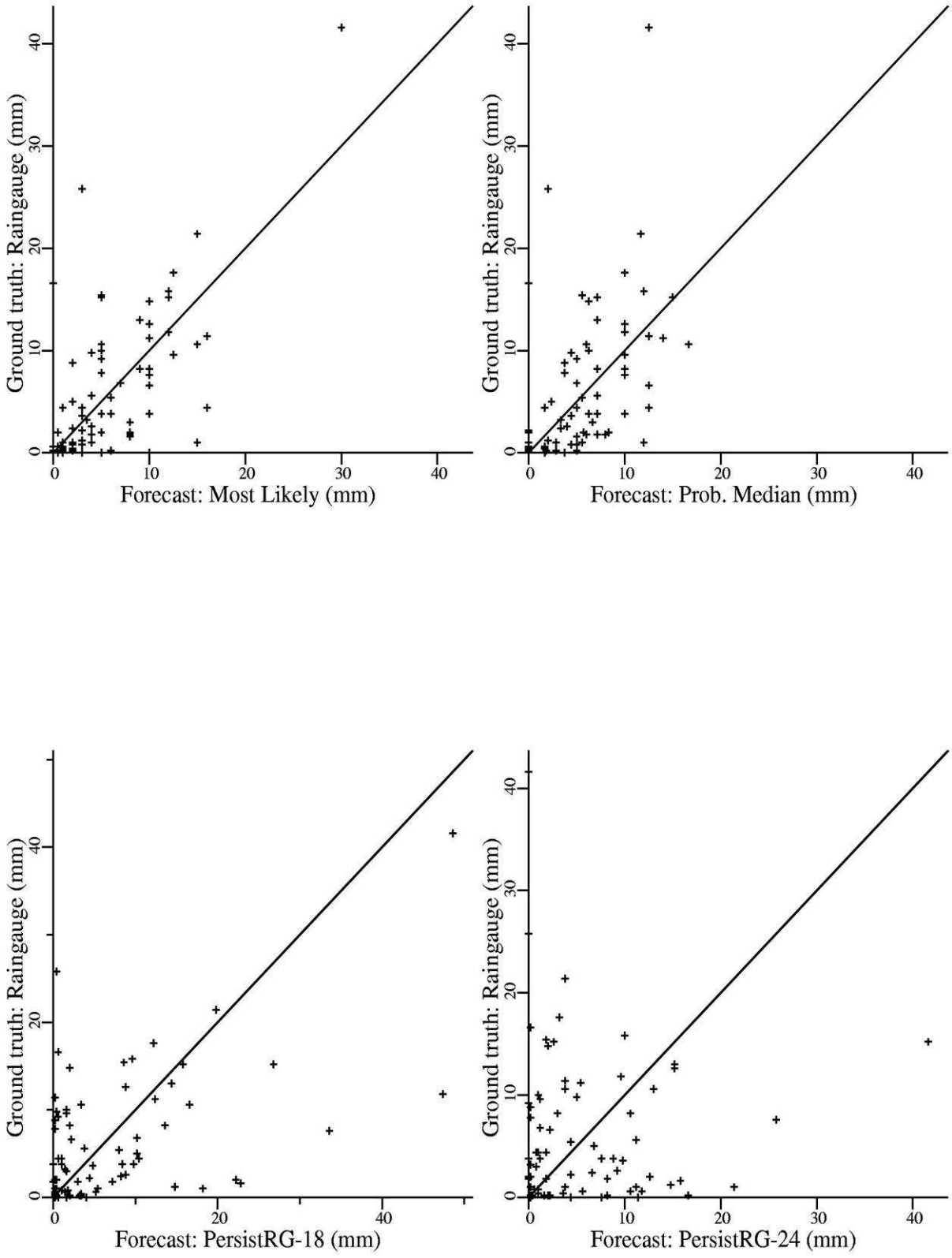


Figure 5.4.2.2 Evening Update forecasts of maximum rainfall amounts. Ground truth from raingauge network. Southeast sub-area of Thames Region.

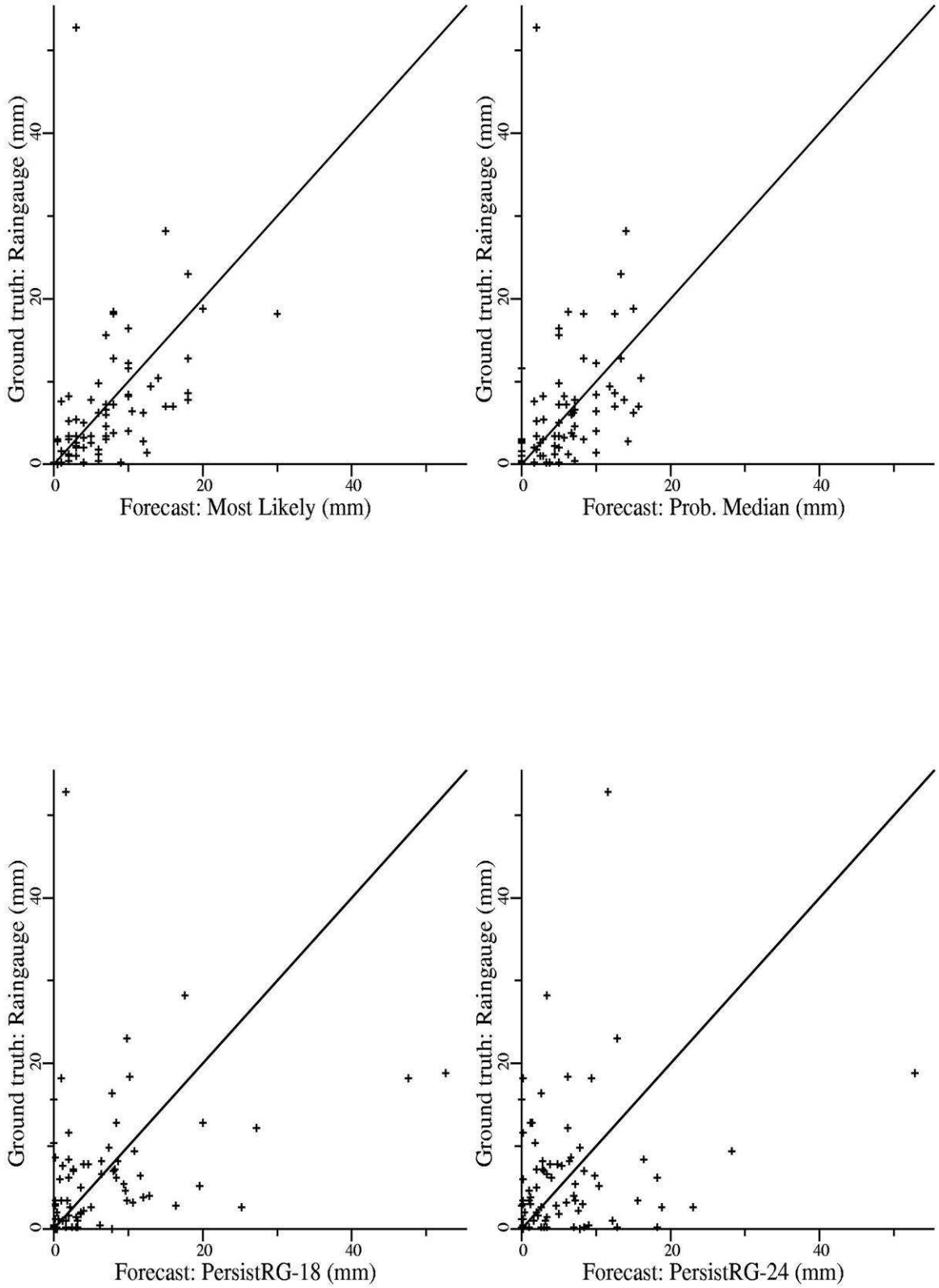


Figure 5.4.2.3 Evening Update forecasts of maximum rainfall amounts. Ground truth from raingauge network. Western sub-area of Thames Region .

Table 5.4.2.2 lists the full set of data for the assessment of forecasts for the North East area of the Agency's Thames Region in the case of the maximum rainfall rates in the 18-hour forecast-period. It will be seen that values for the spatial maximum of rainfall rates obtained from weather radar are usually substantially higher than those obtained from the network of raingauges. (The comparison can be made both for the "outcome" columns and for the "persistence" forecasts.) This may in part be due to the better spatial coverage given by the radar-fields, which would be expected to result in higher estimates of the spatial maximum. There may also be problems arising from the use of data from the Nimrod radar product which was still under development and its access to raingauge data for adjustment was limited. Nonetheless, it is striking that the range of values produced as the "most likely" values in the Evening Update forecasts is rather more similar to that obtained from the raingauge network than from the radar data source. A further point to notice is that, in this example, the outcomes obtained from the radar source are always non-zero: this holds true for each of the three sub-areas. The highest rainfall value derived from the radar source (191.75 mm h^{-1}) was found to occur twice for the Northeast sub-area and once each for the other sub-areas. The multiple occurrence of this value raise some suspicions: while it is not close to the upper-limit of values encompassed by the format used to transmit Nimrod data files, it may be there is an effective upper limit to possible values within the computation procedures being used.

The complete set of scatter plots for the present case study, for the case of forecasts of maximum rainfall rates, is provided in Figures 5.4.2.4 to 5.4.2.9. Once again, these plots indicate that there is not a particularly good correspondence between the operational forecasts and the eventual outcomes and that the performance analyses will not be completely dominated by only one or two particularly bad forecasts. The correspondence between the operational forecasts and the radar-derived ground truth is seen to be particularly poor, with the forecast-values never extending even into the mid-range of the values of the outcomes derived from radar. It can be seen in Figure 5.4.2.5 that the Persistence-forecast derived from already-available radar data can sometimes provide a very good match to the radar-derived outcome. In these cases the maximum rainfall rate occurs very early in the period being forecasted and the value of the forecast arises from the maximum rate being found very late in the period immediately before the forecast period.

Table 5.4.2.2 Example of data for assessment of rainfall forecasts for maximum rainfall rates in an 18-hour time-period: maximum rate in Northeast area of Thames Region (units: mm h^{-1}).

date	-- Evening Update--		----------Persistence----------				---- Outcome----	
	'Most Likely'	Median	18 h Gauge	24 h Gauge	18 h Radar	24 h Radar	Gauge	Radar
25 7 2002 17:00	0.00	0.00	0.80	0.80	7.62	7.62	0.00	17.50
26 7 2002 17:00	0.00	0.00	0.00	0.00	0.25	17.50	0.00	0.09
27 7 2002 17:00	0.00	0.00	0.00	0.00	0.91	0.09	0.00	2.66
28 7 2002 17:00	4.00	0.00	0.00	0.00	3.22	2.66	0.00	2.38
29 7 2002 17:00	15.00	15.00	4.00	0.00	191.75	2.38	28.80	191.75
30 7 2002 17:00	3.00	4.00	54.40	28.80	100.44	191.75	49.60	91.31
31 7 2002 17:00	25.00	15.00	77.60	49.60	79.12	91.31	20.00	79.12
1 8 2002 17:00	5.00	6.00	0.80	20.00	11.03	79.12	0.80	91.75
2 8 2002 17:00	3.00	2.00	0.80	0.80	191.75	191.75	7.20	31.19
3 8 2002 17:00	12.00	8.50	52.80	7.20	60.88	31.19	30.40	109.56
4 8 2002 17:00	30.00	27.50	45.60	30.40	124.78	109.56	4.00	24.34
5 8 2002 17:00	32.00	38.75	26.40	4.00	76.09	24.34	4.00	27.41
6 8 2002 17:00	2.00	1.71	4.00	4.00	7.81	27.41	7.20	13.69
7 8 2002 17:00	15.00	8.00	17.60	7.20	76.09	13.69	42.40	133.94
8 8 2002 17:00	15.00	8.00	32.80	42.40	54.78	133.94	18.40	19.78
9 8 2002 17:00	20.00	20.00	56.80	18.40	79.12	19.78	46.40	170.44
10 8 2002 17:00	6.00	6.00	14.40	46.40	88.28	170.44	19.20	42.62
11 8 2002 17:00	1.00	0.80	0.80	19.20	2.03	42.62	1.60	5.12
4 9 2002 17:00	0.00	0.00	0.80	0.00	0.09	0.72	0.80	82.44
5 9 2002 17:00	3.00	0.00	0.80	0.80	82.44	82.44	6.40	7.94
6 9 2002 17:00	10.00	10.00	6.40	6.40	20.59	7.94	26.40	34.06
7 9 2002 17:00	3.00	2.00	26.40	26.40	34.06	34.06	1.60	9.53
8 9 2002 17:00	20.00	13.33	8.80	1.60	34.25	9.53	12.00	29.88
9 9 2002 17:00	10.00	10.00	41.60	12.00	68.56	29.88	12.00	22.16
10 9 2002 17:00	0.00	0.00	0.80	12.00	1.03	22.16	0.00	0.44
8 10 2002 17:00	2.00	0.00	0.00	0.00	2.72	3.84	0.80	0.84
9 10 2002 17:00	0.00	0.00	0.80	0.80	0.09	0.84	1.60	2.91
10 10 2002 17:00	0.00	0.00	1.60	1.60	5.69	2.91	1.60	1.25
11 10 2002 17:00	4.00	5.00	1.60	1.60	5.53	1.25	20.80	11.41
12 10 2002 17:00	0.00	0.00	20.80	20.80	11.41	11.41	0.80	2.47
13 10 2002 17:00	5.00	6.00	7.20	0.80	9.62	2.47	5.60	19.31
14 10 2002 17:00	8.00	8.50	17.60	5.60	6.47	19.31	9.60	20.09
15 10 2002 17:00	8.00	8.00	25.60	9.60	35.25	20.09	2.40	9.53
16 10 2002 17:00	1.50	1.00	2.40	2.40	5.47	9.53	0.80	1.97
17 10 2002 17:00	0.00	0.00	0.80	0.80	0.69	1.97	2.40	27.94
18 10 2002 17:00	0.00	0.00	2.40	2.40	27.94	27.94	0.80	0.28
19 10 2002 17:00	2.00	2.86	0.80	0.80	0.00	0.28	0.80	2.09
20 10 2002 17:00	5.00	8.50	2.40	0.80	4.50	2.09	8.80	10.81
21 10 2002 17:00	6.00	8.50	30.40	8.80	34.47	10.81	14.40	29.31
22 10 2002 17:00	8.00	11.67	17.60	14.40	36.41	29.31	5.60	9.28
23 10 2002 17:00	3.00	0.00	16.00	5.60	39.00	9.28	0.80	0.97
24 10 2002 17:00	1.00	1.60	0.80	0.80	1.84	0.97	11.20	18.44
25 10 2002 17:00	24.00	27.50	17.60	11.20	102.88	18.44	12.00	27.09
26 10 2002 17:00	6.00	7.00	12.00	12.00	1.44	27.09	11.20	38.53
27 10 2002 18:00	0.50	0.00	11.20	11.20	38.53	38.53	1.60	2.72
28 10 2002 18:00	8.00	5.50	0.80	1.60	1.72	2.72	0.00	2.44
29 10 2002 18:00	6.00	7.60	5.60	0.00	5.00	2.44	1.60	5.56
30 10 2002 18:00	6.00	5.20	8.00	1.60	12.50	5.56	3.20	3.97
31 10 2002 18:00	0.50	0.00	3.20	3.20	5.59	3.97	0.80	3.34
1 11 2002 18:00	5.00	5.50	6.40	0.80	11.22	3.34	2.40	27.81
2 11 2002 18:00	8.00	8.80	4.00	2.40	27.81	27.81	6.40	62.91
3 11 2002 18:00	4.00	0.00	5.60	6.40	41.16	62.91	1.60	16.03
4 11 2002 18:00	2.00	1.00	0.80	1.60	2.88	16.03	0.00	0.09

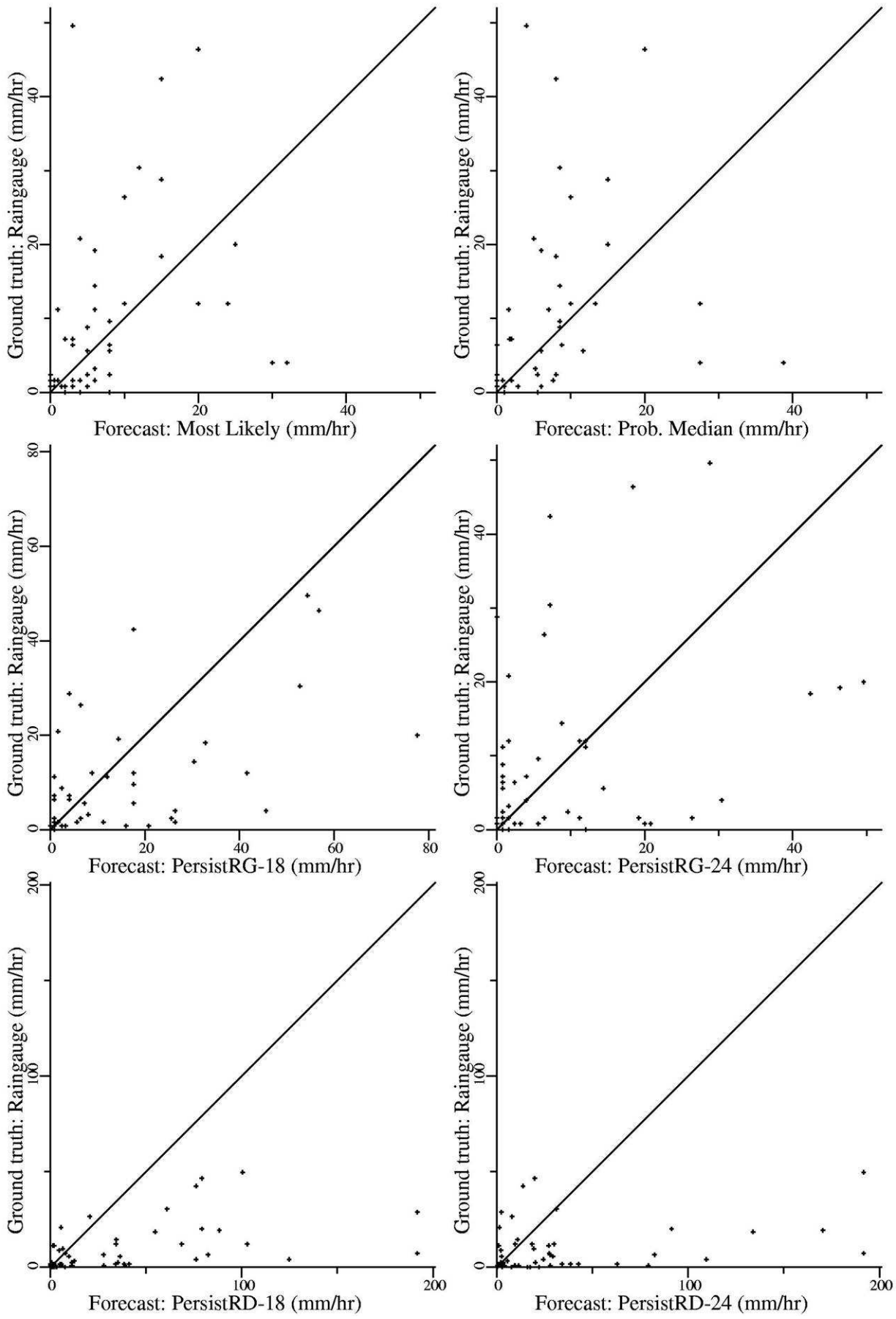


Figure 5.4.2.4 Evening Update forecasts of maximum Rainfall Rate. Ground truth from raingauge network. Northeast sub-area of Thames. Region

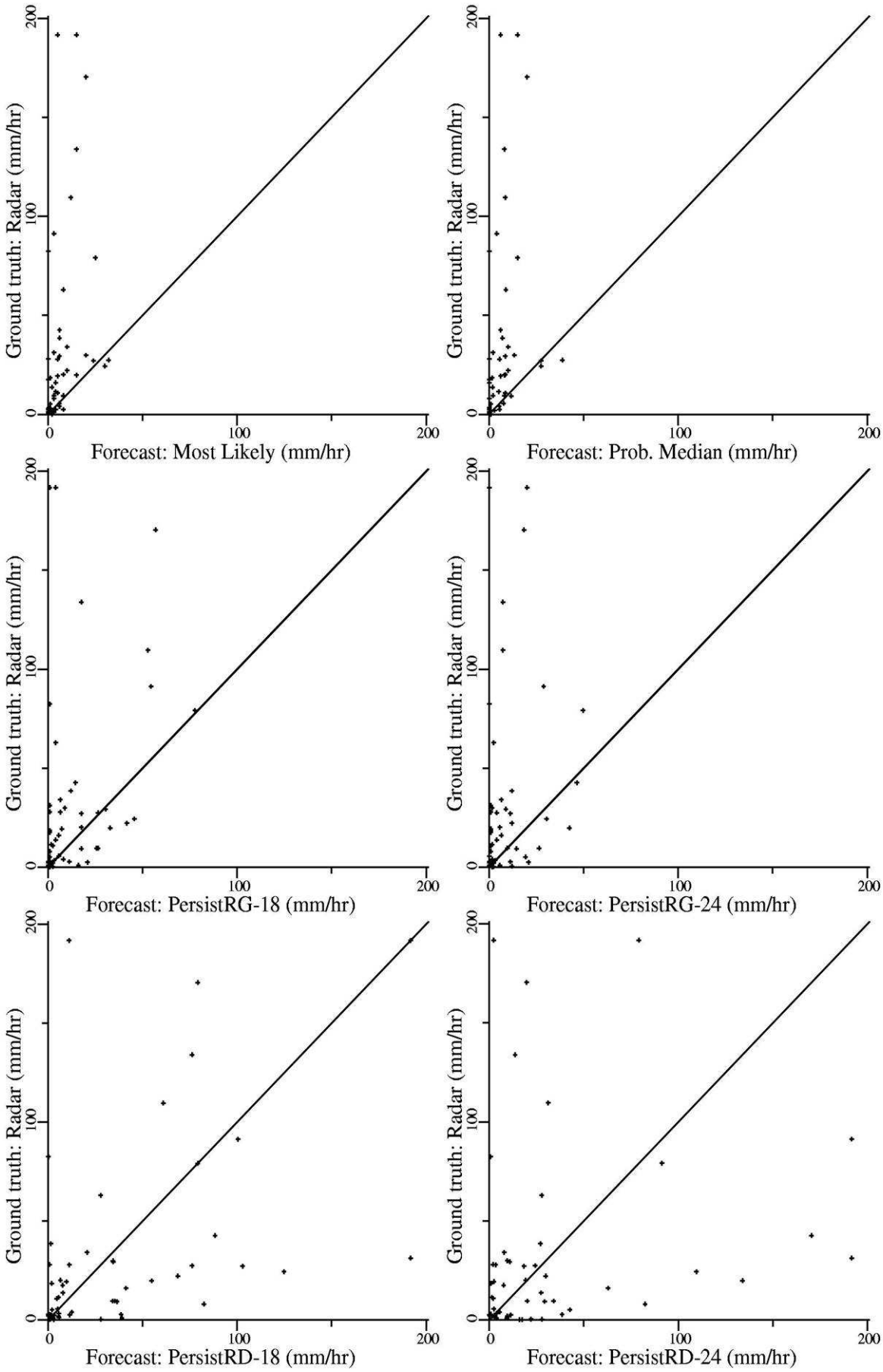


Figure 5.4.2.5 Evening Update forecasts of maximum Rainfall Rate. Ground truth from Nimrod QC Radar. Northeast sub-area of Thames Region.

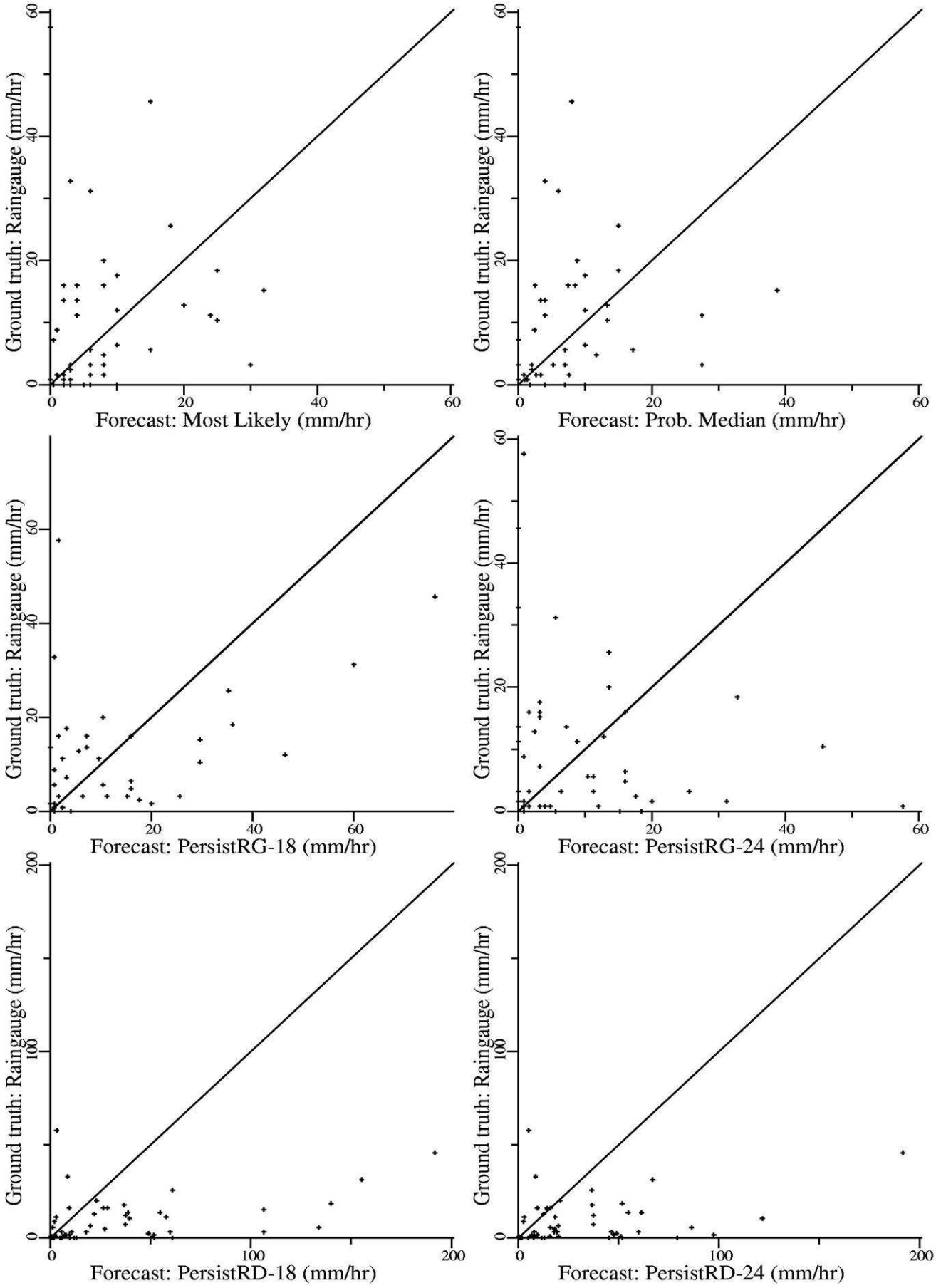


Figure 5.4.2.6 Evening Update forecasts of maximum Rainfall Rate. Ground truth from raingauge network. Southeast sub-area of Thames Region.

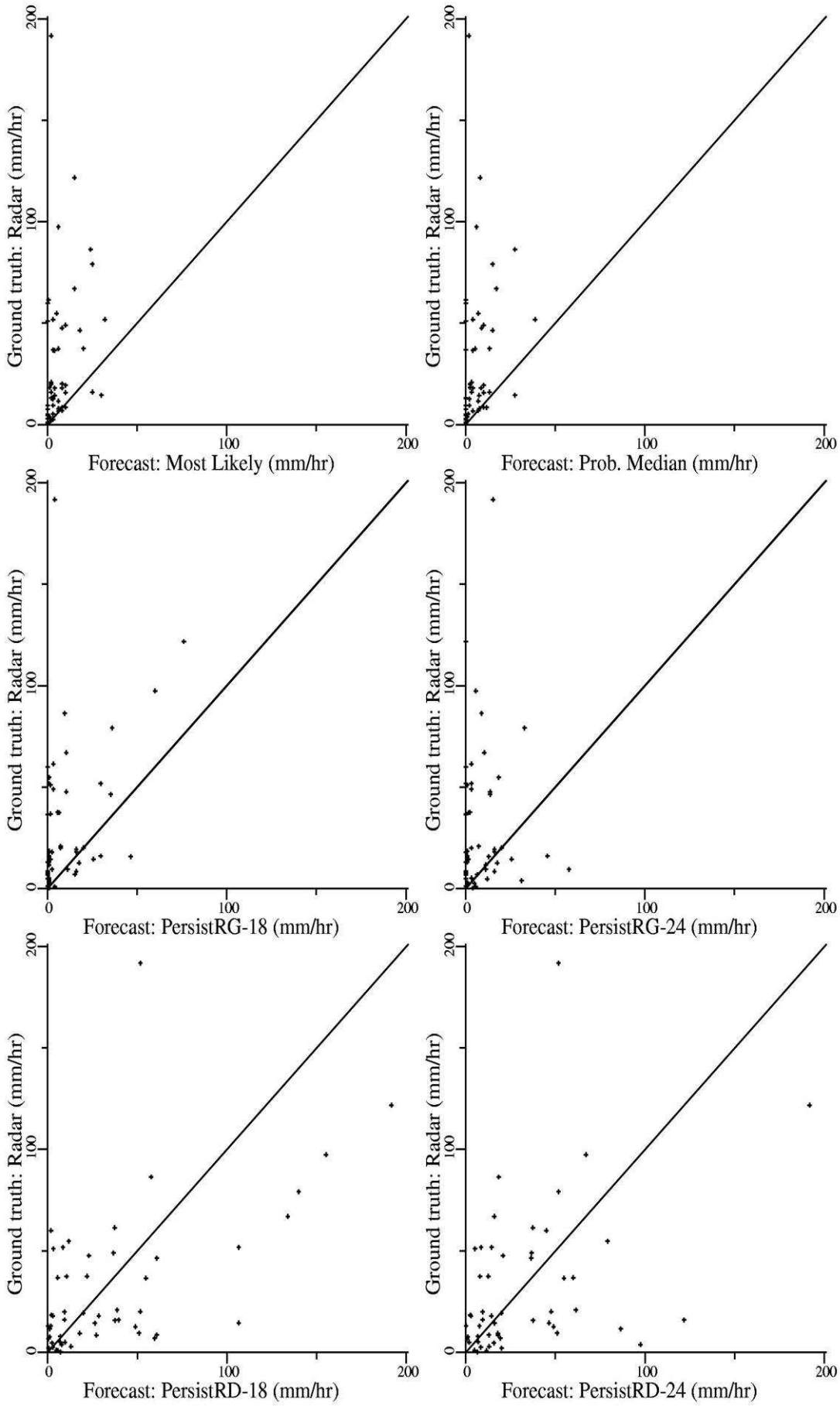


Figure 5.4.2.7 Evening Update forecasts of maximum Rainfall Rate. Ground truth from Nimrod QC Radar. Southeast sub-area of Thames Region.

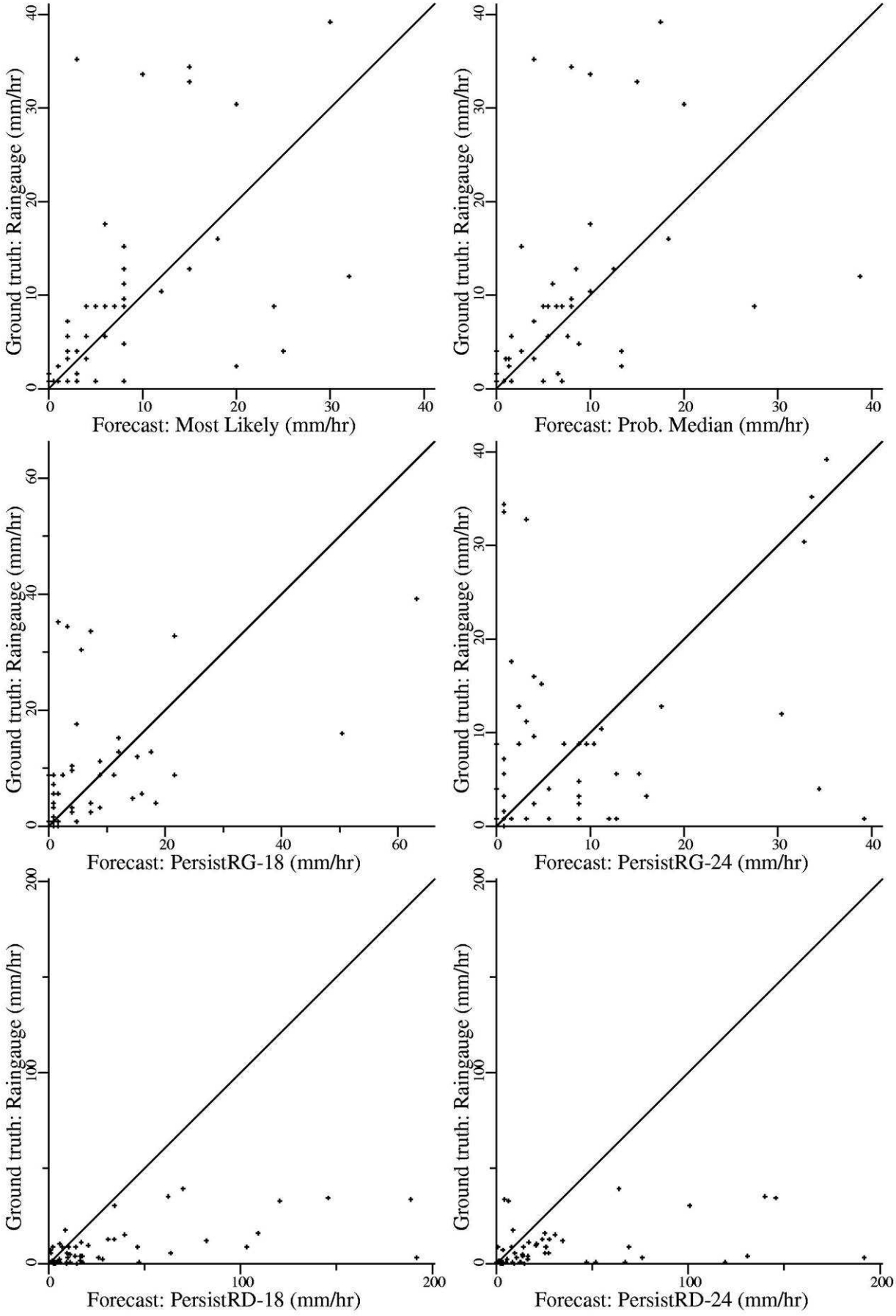


Figure 5.4.2.8 Evening Update forecasts of maximum Rainfall Rate. Ground truth from raingauge network. Western sub-area of Thames Region.

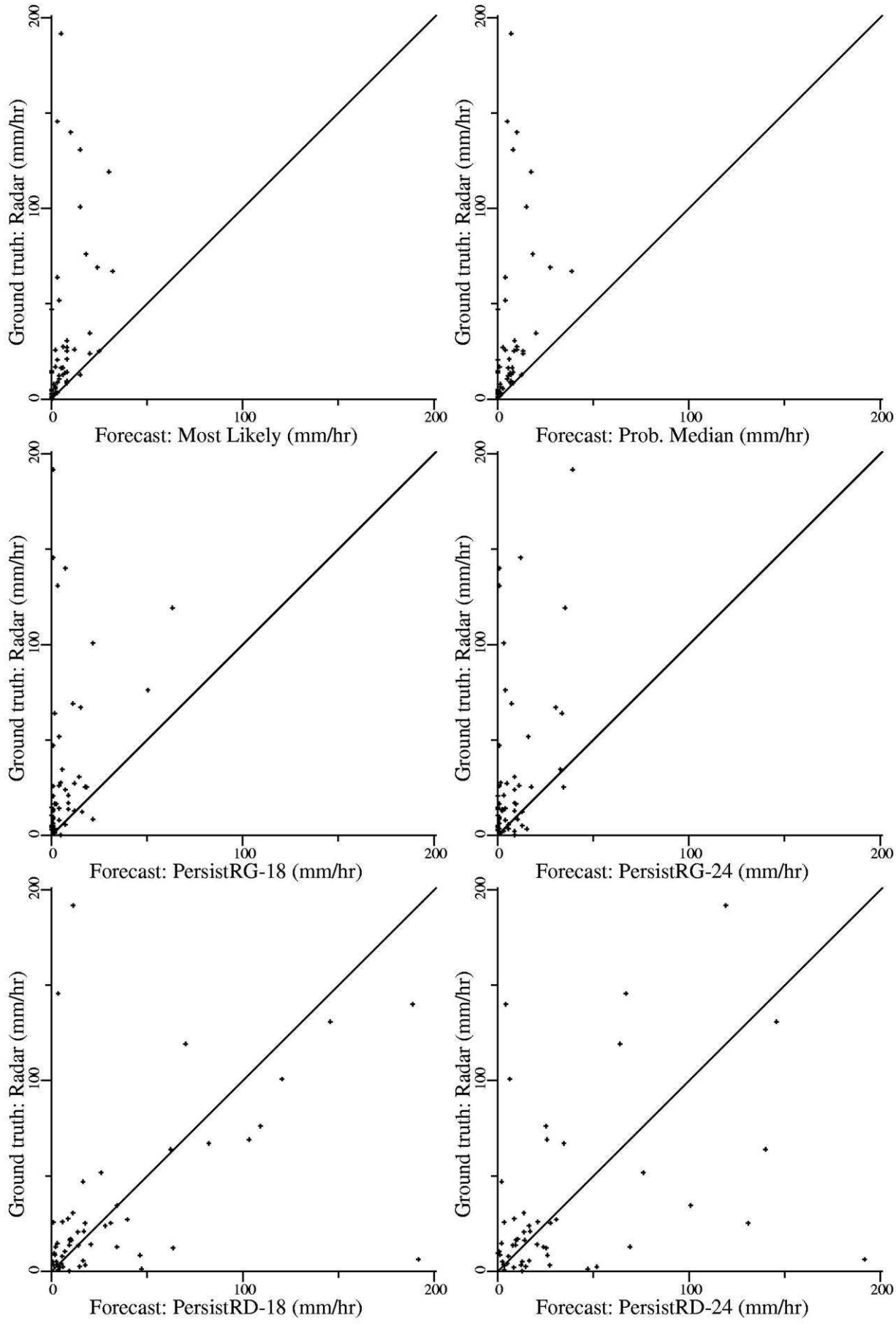


Figure 5.4.2.9 Evening Update forecasts of maximum Rainfall Rate. Ground truth from Nimrod QC Radar. Western sub-area of Thames Region.

5.4.3 Assessment of Single-valued Forecasts of Accumulations

5.4.3.1 Assessment of forecast amounts

Section 2.2.3 has outlined a number of measures of forecast performance appropriate for single-valued forecasts of rainfall amounts. Several of these have been evaluated for the Evening Update forecasts for Thames Region, and the results are presented in Tables 5.4.3.1.1-6.

Table 5.4.3.1.1 shows the basic assessment measures for the size of forecast errors for rainfall amounts, evaluated for the 3 sub-areas of the Thames Region. Results are given for the 6 types of forecasts listed in Table 5.4.1.1 and, in addition, the result is given for the best performance measure obtainable by a constant-value forecast (rows labelled "$Const_{best}$"). Table 5.4.3.1.2 shows the corresponding R^2 (efficiency) measures: these effectively compare the values of the performance measures shown in Table 5.4.3.1.1 with the best performance measure achievable by a constant-value forecast.

The results in Tables 5.4.3.1.1 and 5.4.3.1.2 illustrate that the performance measures for the different sources of forecasts have the expected ranking, with the forecasts from the Evening Updates being better than both persistence forecasts and constant-value forecasts. As might be expected, 18-hour-delayed persistence forecasts are better than the 24-hour-delayed persistence forecasts. However, the R^2 (efficiency) measures for the persistence forecasts are usually negative, indicating that a better forecast performance can be achieved by selecting a suitable constant-value forecast. While there is little difference in performance between the two forecasts obtained from the Evening Updates, the values taken directly from the forecasts, given by the 'most likely' values, are usually better than the forecast derived as the median of the probability forecasts. As discussed earlier, this may be partly due to the inaccuracy involved in expressing the probability forecast in the form of a simple table with limited resolution.

Table 5.4.3.1.3 shows details of the bias contained in the various forecast sources. Here the usual statistical practice is followed of defining the direction in which an "error" is measured as being positive if the outcome is larger than the forecast, and hence the bias being negative means that the forecast tends to be larger than the actual outcome. Overall it seems that the Evening Update forecasts give values which are slightly too large, with the forecasts derived as the median of the probability forecasts tending to be smaller than the 'most likely' values. Table 5.4.3.1.4 shows some statistics for the rainfall amounts which give more details of the typical amounts obtained for the actual outcomes and for the forecasts. This table shows that the variation between the sub-areas of the typical amounts observed for the outcome is reasonably closely followed by the variation of the typical amounts being forecasted in the Evening Updates. The forecast values have standard-deviations rather lower than the actual outcomes, a feature which would be expected in most forecasting situations.

Table 5.4.3.1.5 gives values for correlation and regression coefficients for linear relationships between outcomes and forecasts of rainfall and log-rainfall. The values here indicate that modest improvements may be obtained to the raw forecasts already considered by forming a simple adjustment of the form:

$$f_i^* = \mu_o + \beta(f_i - \mu_f).$$

Here, f_i^* is the new forecast value constructed from the raw value f_i for occasion i, β is the regression coefficient and μ_o and μ_f are the means of the outcome and raw forecast values. The extent of potential improvement can be judged by comparing the square of the correlation coefficient with the value, in Table 5.4.3.1.2, for R^2 for Root Mean Square Error. For example, the R^2 for the "Most Likely" forecast in the NE sub-area might be increased from 0.16 to 0.21 for a simple scaling of the rainfall amount, or, in the log-space, from 0.30 to 0.44 for an adjustment to the logarithm of rainfall amounts. Such potential adjustments are often not pursued because one effect of the adjustment is that forecasts on occasions when the raw forecast is zero will no longer be zero: in the example used above, for a simple adjustment of the "Most Likely" forecast, the smallest value forecasted would be 1.75mm (=5.08-0.64×5.20). In addition the parameters used in the adjustment are themselves values estimated from only limited data and the effect of carrying forward such estimated adjustment parameters is open to concern.

The above analysis of performance has been the traditional one where standard measures of forecast performance are evaluated separately for each forecast source and then compared. As discussed in Section 4.3, it is possible to do a rather more detailed analysis and to determine whether the evidence provided by the test dataset is sufficient to distinguish between the performance of different forecast sources, bearing in mind the sampling variability of the forecast performance statistics and the statistical dependences between them. Table 5.4.3.1.6 relates directly to this question. Taking the 'most likely' forecast ("Most Likely") from the Evening Updates as a "base forecast", Table 5.4.3.1.6 considers each of the other forecast sources in turn and asks how much evidence there is that the "base forecast" is better than the candidate forecast. The values given are the standardised differences discussed earlier in Section 4.3.3, and positive values indicate that the "base forecast" has a smaller size of error, as measured by the performance statistic, than the candidate. If the candidate forecast produces smaller errors, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the long-run performance measures for the two forecast sources will turn out to be in the order indicated. For the purposes here, a standardised difference outside the range ± 2 units indicates fairly strong evidence that one forecast source really is better than another.

The results in Table 5.4.3.1.6 reflect those in Table 5.4.3.1.1, in that the comparisons which favour one forecast source over another are the same. However, Table 5.4.3.1.6 provides extra information. For example, it shows that there is only weak evidence that the 'most likely' values in the Evening Update provide better forecasts than the median-values derived from the probability forecasts. There is fairly strong evidence, for all the performance measures, that the "Most Likely" forecast is better than a persistence based on the 18-hours immediately before the start of the forecast period. (It should be recalled that the Evening Updates are typically issued 2 hours before the start of the forecast period.)

It appears from Table 5.4.3.1.6 that the performance measures based on errors in the logarithms of rainfall amounts are able to provide stronger evidence in the comparison of forecast sources than do those based on ordinary errors. Similarly, the Root Mean Square Error performance measure appears to provide weaker evidence for differences than does the Mean Absolute Error. These appearances may be misleading: the effect is related to certain of the performance measures being more or less sensitive to errors in the forecast when rainfall outcomes or forecasts are large. Some performance measures emphasise these (squared-error criteria compared with absolute-error criteria), or discount these (those based on errors of logarithms compared with those using ordinary errors), and hence may be more or less sensitive to individual outcomes. Less sensitive performance measures may be able to yield stronger evidence for differences between forecast sources, but they may not adequately reflect the uses to which forecasts are put. It is arguable that performance of forecasts when rainfall amounts are high should certainly not be discounted against performance in low rainfall conditions, since it is exactly those high-rainfall occasions when the forecasts are most important.

Table 5.4.3.1.1 Raw assessment measures for Evening Update forecasts in the Thames Region. (Rainfall Totals)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error (mm)	Most Likely	3.73	3.50	3.83
	Prob. Median	3.88	3.61	3.94
	$Persist_{RG,18}$	5.26	5.27	5.46
	$Persist_{RG,24}$	5.60	6.44	6.24
	$Const_{0mm}$	5.08	5.59	6.00
	$Const_{5mm}$	4.76	5.12	4.99
	$Const_{best}$	4.42	4.83	4.81
Root Mean Square Error (mm)	Most Likely	6.72	5.32	7.20
	Prob. Median	6.80	5.96	7.38
	$Persist_{RG,18}$	9.25	8.42	9.74
	$Persist_{RG,24}$	9.17	9.58	9.71
	$Const_{0mm}$	8.86	8.99	9.89
	$Const_{5mm}$	7.25	7.07	7.93
	$Const_{best}$	7.25	7.04	7.86
Mean Absolute Error of Log-Rainfall (dimensionless)	Most Likely	1.03	1.00	0.79
	Prob. Median	1.10	0.96	0.98
	$Persist_{RG,18}$	1.30	1.35	1.22
	$Persist_{RG,24}$	1.66	1.71	1.44
	$Const_{0mm}$	2.84	2.95	3.14
	$Const_{5mm}$	1.54	1.57	1.38
	$Const_{best}$	1.46	1.50	1.34
Root Mean Square Error of Log-Rainfall (dimensionless)	Most Likely	1.38	1.46	1.08
	Prob. Median	1.45	1.37	1.43
	$Persist_{RG,18}$	1.76	1.78	1.67
	$Persist_{RG,24}$	2.09	2.23	1.91
	$Const_{0mm}$	3.32	3.44	3.56
	$Const_{5mm}$	2.02	2.01	1.84
	$Const_{best}$	1.71	1.77	1.67

Table 5.4.3.1.2 **R^2 (efficiency) measures for Evening Update forecasts in the Thames Region for each type of assessment measure. (Rainfall Totals)**

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
R^2 for Mean Absolute Error	Most Likely	0.16	0.28	0.20
	Prob. Median	0.12	0.25	0.18
	$Persist_{RG,18}$	-0.19	-0.09	-0.13
	$Persist_{RG,24}$	-0.27	-0.33	-0.30
	$Const_{0mm}$	-0.15	-0.16	-0.25
	$Const_{5mm}$	-0.08	-0.06	0.04
	$Const_{best}$	0.00	0.00	0.00
R^2 for Root Mean Square Error	Most Likely	0.14	0.43	0.16
	Prob. Median	0.12	0.28	0.12
	$Persist_{RG,18}$	-0.63	-0.43	-0.53
	$Persist_{RG,24}$	-0.60	-0.85	-0.53
	$Const_{0mm}$	-0.49	-0.63	-0.58
	$Const_{5mm}$	0.00	-0.01	-0.02
	$Const_{best}$	0.00	0.00	0.00
R^2 for Mean Absolute Error of Log-Rainfall	Most Likely	0.30	0.34	0.41
	Prob. Median	0.25	0.36	0.26
	$Persist_{RG,18}$	0.11	0.10	0.09
	$Persist_{RG,24}$	-0.13	-0.14	-0.08
	$Const_{0mm}$	-0.94	-0.97	-1.35
	$Const_{5mm}$	-0.05	-0.04	-0.03
	$Const_{best}$	0.00	0.00	0.00
R^2 for Root Mean Square Error of Log-Rainfall	Most Likely	0.35	0.32	0.58
	Prob. Median	0.28	0.40	0.27
	$Persist_{RG,18}$	-0.06	-0.01	0.01
	$Persist_{RG,24}$	-0.50	-0.58	-0.31
	$Const_{0mm}$	-2.76	-2.78	-3.53
	$Const_{5mm}$	-0.39	-0.29	-0.21
	$Const_{best}$	0.00	0.00	0.00

Table 5.4.3.1.3 Bias measures for Evening Update forecasts in the Thames Region. (Rainfall Totals)

Bias Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Error (mm)	Most Likely	-0.12	0.06	-0.23
	Prob. Median	0.41	0.55	0.75
	$Persist_{RG,18}$	-0.57	-0.79	0.00
	$Persist_{RG,24}$	0.59	0.26	0.23
	$Const_{0mm}$	5.08	5.59	6.00
	$Const_{5mm}$	0.08	0.59	1.00
Median Error (mm)	Most Likely	-0.80	-0.35	-0.80
	Prob. Median	-0.37	0.00	0.00
	$Persist_{RG,18}$	0.00	0.00	0.00
	$Persist_{RG,24}$	0.00	0.00	0.00
	$Const_{0mm}$	2.90	2.80	3.40
	$Const_{5mm}$	-2.10	-2.20	-1.60
Mean Error of Log-Rainfall (dimensionless)	Most Likely	-0.34	-0.36	-0.25
	Prob. Median	-0.10	-0.06	0.15
	$Persist_{RG,18}$	-0.08	0.06	0.19
	$Persist_{RG,24}$	0.09	0.07	0.07
	$Const_{0mm}$	2.84	2.95	3.14
	$Const_{5mm}$	-1.07	-0.96	-0.77
Median Error of Log-Rainfall (dimensionless)	Most Likely	-0.29	-0.10	-0.19
	Prob. Median	-0.08	0.00	0.00
	$Persist_{RG,18}$	0.00	0.00	0.00
	$Persist_{RG,24}$	0.00	0.00	0.00
	$Const_{0mm}$	3.36	3.33	3.53
	$Const_{5mm}$	-0.55	-0.58	-0.39

Table 5.4.3.1.4 Statistics of forecasts and outcomes for Evening Update forecasts in Thames Region. (Rainfall Totals)

Statistic of Rainfall	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Rainfall (mm)	Outcome	5.08	5.59	6.00
	Most Likely	5.20	5.53	6.23
	Prob. Median	4.67	5.04	5.25
Median Rainfall (mm)	Outcome	2.90	2.80	3.40
	Most Likely	4.00	4.00	5.00
	Prob. Median	4.00	5.00	5.00
Standard Deviation (mm)	Outcome	7.30	7.09	7.91
	Most Likely	5.24	5.24	5.91
	Prob. Median	4.12	4.37	4.82

Table 5.4.3.1.5 Correlation of Evening Update forecasts with outcomes in Thames Region . (Rainfall Totals)

Correlation Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Correlation (dimensionless)	Most Likely	0.46	0.66	0.48
	Prob. Median	0.39	0.54	0.41
	$Persist_{RG,18}$	0.29	0.53	0.35
	$Persist_{RG,24}$	0.15	0.05	0.23
Regression Coefficient (dimensionless)	Most Likely	0.64	0.89	0.64
	Prob. Median	0.70	0.88	0.67
	$Persist_{RG,18}$	0.26	0.39	0.30
	$Persist_{RG,24}$	0.16	0.05	0.23
Correlation of Log-Rainfall (dimensionless)	Most Likely	0.67	0.64	0.79
	Prob. Median	0.67	0.71	0.69
	$Persist_{RG,18}$	0.48	0.51	0.53
	$Persist_{RG,24}$	0.23	0.22	0.36
Regression Coefficient of Log-Rainfall (dimensionless)	Most Likely	0.73	0.75	0.84
	Prob. Median	0.63	0.68	0.61
	$Persist_{RG,18}$	0.47	0.50	0.51
	$Persist_{RG,24}$	0.24	0.22	0.35

Table 5.4.3.1.6 Comparison of forecast sources: Standardised differences for assessment measures for Evening Update forecasts in the Thames Region. (Rainfall Totals)
(In this Table, the base forecast is "Most Likely")

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error	Prob. Median	0.73	0.35	0.44
	$Persist_{RG,18}$	2.06	2.62	2.48
	$Persist_{RG,24}$	2.83	4.37	3.33
	$Const_{0mm}$	2.67	3.22	4.04
	$Const_{5mm}$	2.31	3.28	2.65
Root Mean Square Error	Prob. Median	0.43	0.79	0.74
	$Persist_{RG,18}$	1.64	2.24	2.33
	$Persist_{RG,24}$	1.82	2.86	2.09
	$Const_{0mm}$	2.96	2.47	3.63
	$Const_{5mm}$	1.13	1.38	1.45
Mean Absolute Error of Log-Rainfall	Prob. Median	0.79	-0.33	2.04
	$Persist_{RG,18}$	1.82	2.28	3.15
	$Persist_{RG,24}$	3.75	4.14	4.40
	$Const_{0mm}$	7.45	7.45	10.74
	$Const_{5mm}$	3.61	3.98	4.22
Root Mean Square Error of Log-Rainfall	Prob. Median	0.57	-0.58	2.28
	$Persist_{RG,18}$	1.98	1.66	2.86
	$Persist_{RG,24}$	3.64	3.37	3.89
	$Const_{0mm}$	7.70	7.68	10.35
	$Const_{5mm}$	4.06	3.17	4.15

5.4.3.2 Assessment of category-forecasts

In addition to dealing with forecasts of rainfall amounts, Section 2.2.4 has outlined a number of measures of forecast performance appropriate for use where forecasts are in the form of simple statements as to whether or not a certain threshold will be exceeded.. The forecasts provided by the Evening Updates can readily be converted to be of this form and, since a number of different thresholds of rainfall amounts can be selected, they provide a useful means of assessing the underlying forecasts' ability to distinguish between zero- and non-zero rainfall conditions and moderate and high-rainfall conditions. For the present study a number of thresholds for rainfall amounts have been chosen which are perhaps unrealistic for practical use, but they illustrate the problems involved in attempting to specify performance measures for categorical forecasts in circumstances where the numbers of cases are limited.

Tables 5.4.3.2.1 to 5.4.3.2.4 show results for a collection of performance measures for analyses using thresholds of 0, 4, 8 and 12mm for the maximum 18-hour rainfall accumulations in each of the 3 sub-areas of the Thames Region. Results are given for the 6 types of forecasts listed in Table 5.4.1.1 and, in addition, results are given for the values of the performance measures if forecasts of exceedences and non-exceedences of the threshold were made at random with the same rate of occurrence as found for the outcomes across all of the test occasions included in this study. The results for this type of forecast are listed against the name "Climatology": they provide a point of comparison for the candidate forecasts since a good forecast should do much better than the type of random forecast represented by "Climatology". For completeness, results are given for a second type of random forecast: these appear in parentheses after the actual values for the performance measure. In these cases, the random forecasts have a rate of forecasting threshold-exceedence equal to that observed for the actual forecasts.

The types of performance measures available for categorical forecasts fall naturally into two groups, and each table is divided in two corresponding parts. In the first group are the ordinary score statistics in which the performance measures are defined fairly directly in terms of the rates of occurrences of success or failure of the forecasts: these are listed in part **a** of each Table. The second group includes more refined measures in which the forecast performance is measured relative to what could be achieved by random forecasts of the two types outlined above: these are listed in part **b** of each Table.

In constructing the tables of results for performance measures of categorical forecasts, there are many cases where the values cannot be calculated because of the need to divide by zero: in these cases the results are represented by an asterisk (*). This rule has been applied even in cases where the standard formula formally gives 0/0 and where there is a potential to create a meaningful numerical value by re-expressing the formula in an alternative way.

The ordinary performance scores for a threshold of 0 mm (Table 5.4.3.2.1a) suggest that the Evening Update forecasts are not substantially better than random forecasts in forecasting whether or not there will be rain: this impression is contradicted by the relative score measures (Table 5.4.3.2.1b) which suggest that even the persistence forecasts provide a worthwhile improvement over random forecasts. Overall, the

performance measures do not provide any clear ordering among the Evening Update and persistence forecasts. For example, for the North East sub-area, the persistence forecast taken from the immediately preceding 18 hours is preferred over the "Most Likely" Evening Update forecast according to the Heidke Skill Score, the Equitable Threat Score, the Likelihood Ratio criterion for values not exceeding the threshold and the Odds Ratio. The reverse is true for the Kuipers Skill Score and the Likelihood Ratio criterion for values which do exceed the threshold. If a comparison is attempted between the performance of the 'most likely' values and the median of the probability forecasts from the Evening Updates, a similar disparity of results occurs. The apparent preference for the various forecasts varies between the 3 sub-areas and this suggests that the performance scores are not well-determined by the amount of data available, at least for this threshold. This underlines the need to develop a means of quantifying and taking into account the sampling variability inherent in the performance scores for categorical forecasts.

When similar analyses are made for the cases of thresholds at 4 mm and 8 mm (Tables 5.4.3.2.2 and 5.4.3.2.3), there is a much clearer consensus (although not unanimity) of the candidate forecasts, across all of the relative performance scores and all of the sub-areas. The analyses indicate that the main four contending forecasts should be put in the order: 'Most Likely', 'Probability Forecast Median', 'Persistence based on the previous 18 hours' and 'Persistence starting 24 hours previously'. When the threshold is raised to 12 mm (Table 5.4.3.2.4), the 'Persistence based on the previous 18 hours' is preferred across all of the relative performance measures and all the sub-areas.

Table 5.4.3.2.1a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 0.0mm

Ordinary Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
H Hit Rate	Most Likely	0.87 (0.79)	0.87 (0.78)	0.93 (0.82)
	Prob. Median	0.79 (0.70)	0.83 (0.68)	0.80 (0.70)
	$Persist_{RG,18}$	0.91(0.85)	0.93 (0.79)	0.89 (0.85)
	$Persist_{RG,24}$	0.88 (0.82)	0.88 (0.75)	0.91 (0.83)
	$Const_{0mm}$	0.10 (0.10)	0.13 (0.13)	0.07 (0.07)
	$Const_{5mm}$	0.90 (0.90)	0.87 (0.87)	0.93 (0.93)
	Climatology	0.82	0.82	0.86
CSI Critical Success Index	Most Likely	0.86 (0.79)	0.86 (0.77)	0.92 (0.82)
	Prob. Median	0.78 (0.69)	0.81 (0.67)	0.79 (0.69)
	$Persist_{RG,18}$	0.91 (0.85)	0.92 (0.78)	0.89 (0.85)
	$Persist_{RG,24}$	0.87 (0.82)	0.87 (0.74)	0.91 (0.83)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.90 (0.90)	0.87 (0.87)	0.93 (0.93)
	Climatology	0.82	0.76	0.86
FAR False Alarm Rate	Most Likely	0.06 (0.10)	0.08 (0.13)	0.01 (0.07)
	Prob. Median	0.03 (0.10)	0.03 (0.13)	0.00 (0.07)
	$Persist_{RG,18}$	0.06 (0.10)	0.05 (0.13)	0.05 (0.07)
	$Persist_{RG,24}$	0.07 (0.10)	0.06 (0.13)	0.03 (0.07)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	0.10 (0.10)	0.13 (0.13)	0.07 (0.07)
	Climatology	0.10	0.13	0.07
POD Probability of Detection	Most Likely	0.91 (0.87)	0.93 (0.88)	0.93 (0.88)
	Prob. Median	0.80 (0.74)	0.83 (0.74)	0.79 (0.73)
	$Persist_{RG,18}$	0.97 (0.94)	0.97 (0.89)	0.93 (0.91)
	$Persist_{RG,24}$	0.93 (0.90)	0.92 (0.84)	0.93 (0.89)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	0.90	0.87	0.93
B Bias Ratio	Most Likely	0.96 (0.96)	1.01 (1.01)	0.95 (0.95)
	Prob. Median	0.82 (0.82)	0.86 (0.86)	0.79 (0.79)
	$Persist_{RG,18}$	1.04 (1.04)	1.03 (1.03)	0.99 (0.99)
	$Persist_{RG,24}$	1.00 (1.00)	0.97 (0.97)	0.96 (0.96)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	1.11 (1.11)	1.15 (1.15)	1.08 (1.08)
	Climatology	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.1b **Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 0.0mm**

Relative Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
HSS Heidke Skill Score	Most Likely	0.35 (0.00)	0.40 (0.00)	0.59 (0.00)
	Prob. Median	0.32 (0.00)	0.47 (0.00)	0.35 (0.00)
	$Persist_{RG,18}$	0.42 (0.00)	0.66 (0.00)	0.25 (0.00)
	$Persist_{RG,24}$	0.31 (0.00)	0.51 (0.00)	0.49 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.41 (0.00)	0.38 (0.00)	0.77 (0.00)
	Prob. Median	0.55 (0.00)	0.65 (0.00)	0.79 (0.00)
	$Persist_{RG,18}$	0.35 (0.00)	0.61 (0.00)	0.27 (0.00)
	$Persist_{RG,24}$	0.31 (0.00)	0.55 (0.00)	0.60 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.21 (0.00)	0.25 (0.00)	0.42 (0.00)
	Prob. Median	0.19 (0.00)	0.31 (0.00)	0.22 (0.00)
	$Persist_{RG,18}$	0.26 (0.00)	0.49 (0.00)	0.14 (0.00)
	$Persist_{RG,24}$	0.16 (0.00)	0.34 (0.00)	0.32 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	1.81 (1.00)	1.70 (1.00)	5.61 (1.00)
	Prob. Median	3.19 (1.00)	4.57 (1.00)	* (1.00)
	$Persist_{RG,18}$	1.56 (1.00)	2.67 (1.00)	1.40 (1.00)
	$Persist_{RG,24}$	1.59 (1.00)	2.52 (1.00)	2.80 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	5.29 (1.00)	6.45 (1.00)	12.67 (1.00)
	Prob. Median	3.70 (1.00)	4.84 (1.00)	4.75 (1.00)
	$Persist_{RG,18}$	13.88 (1.00)	22.59 (1.00)	5.07 (1.00)
	$Persist_{RG,24}$	5.55 (1.00)	7.53 (1.00)	10.13 (1.00)
	$Const_{0mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00
θ Odds Ratio	Most Likely	9.57 (1.00)	11.00 (1.00)	71.00 (1.00)
	Prob. Median	11.80 (1.00)	22.12 (1.00)	* (1.00)
	$Persist_{RG,18}$	21.60 (1.00)	60.38 (1.00)	7.10 (1.00)
	$Persist_{RG,24}$	8.28 (1.00)	18.96(1.00)	28.40 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.2a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 4.0mm

Ordinary Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
H Hit Rate	Most Likely	0.82 (0.51)	0.73 (0.50)	0.77 (0.50)
	Prob. Median	0.76 (0.50)	0.72 (0.49)	0.74 (0.49)
	$Persist_{RG,18}$	0.68(0.52)	0.65 (0.52)	0.71 (0.51)
	$Persist_{RG,24}$	0.59 (0.52)	0.50 (0.52)	0.57 (0.51)
	$Const_{0mm}$	0.59 (0.59)	0.59 (0.59)	0.56 (0.56)
	$Const_{5mm}$	0.41 (0.41)	0.41 (0.41)	0.44 (0.44)
	Climatology	0.51	0.51	0.51
CSI Critical Success Index	Most Likely	0.65 (0.28)	0.54 (0.29)	0.61 (0.31)
	Prob. Median	0.57 (0.29)	0.55 (0.31)	0.59 (0.32)
	$Persist_{RG,18}$	0.43 (0.25)	0.36 (0.25)	0.48 (0.26)
	$Persist_{RG,24}$	0.32 (0.25)	0.24 (0.26)	0.34 (0.28)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.41 (0.41)	0.41 (0.41)	0.44 (0.44)
	Climatology	0.26	0.26	0.86
FAR False Alarm Rate	Most Likely	0.24 (0.59)	0.35 (0.59)	0.30 (0.56)
	Prob. Median	0.32 (0.59)	0.38 (0.59)	0.33 (0.56)
	$Persist_{RG,18}$	0.38 (0.59)	0.42 (0.59)	0.31 (0.56)
	$Persist_{RG,24}$	0.50 (0.59)	0.61 (0.59)	0.49 (0.56)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	0.59 (0.59)	0.59 (0.59)	0.56 (0.56)
	Climatology	0.59	0.59	0.56
POD Probability of Detection	Most Likely	0.82 (0.45)	0.76 (0.88)	0.83 (0.52)
	Prob. Median	0.79 (0.49)	0.82 (0.74)	0.83 (0.55)
	$Persist_{RG,18}$	0.59 (0.39)	0.53 (0.89)	0.61 (0.39)
	$Persist_{RG,24}$	0.47 (0.39)	0.38 (0.84)	0.50 (0.43)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	0.41	0.41	0.44
B Bias Ratio	Most Likely	1.09 (1.09)	1.18 (1.18)	1.19 (1.19)
	Prob. Median	1.18 (1.18)	1.32 (0.32)	1.25 (1.25)
	$Persist_{RG,18}$	0.94 (0.94)	0.91 (0.91)	0.89 (0.89)
	$Persist_{RG,24}$	0.94 (0.94)	0.97 (0.97)	0.97 (0.97)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	2.41 (2.41)	2.41 (2.41)	2.28 (2.28)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.2b **Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 4.0mm**

Relative Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
HSS Heidke Skill Score	Most Likely	0.63 (0.00)	0.46 (0.00)	0.54 (0.00)
	Prob. Median	0.51 (0.00)	0.45 (0.00)	0.49 (0.00)
	$Persist_{RG,18}$	0.34 (0.00)	0.26 (0.00)	0.40 (0.00)
	$Persist_{RG,24}$	0.14 (0.00)	-0.03 (0.00)	0.13 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.64 (0.00)	0.47 (0.00)	0.55 (0.00)
	Prob. Median	0.52 (0.00)	0.47 (0.00)	0.51 (0.00)
	$Persist_{RG,18}$	0.34 (0.00)	0.26 (0.00)	0.39 (0.00)
	$Persist_{RG,24}$	0.14 (0.00)	-0.03 (0.00)	0.13 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.46 (0.00)	0.30 (0.00)	0.37 (0.00)
	Prob. Median	0.34 (0.00)	0.29 (0.00)	0.33 (0.00)
	$Persist_{RG,18}$	0.21 (0.00)	0.15 (0.00)	0.25 (0.00)
	$Persist_{RG,24}$	0.07 (0.00)	-0.02 (0.00)	0.07 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	4.39 (1.00)	2.62 (1.00)	2.95 (1.00)
	Prob. Median	2.93 (1.00)	2.33 (1.00)	2.56 (1.00)
	$Persist_{RG,18}$	2.35 (1.00)	1.95 (1.00)	2.81 (1.00)
	$Persist_{RG,24}$	1.41 (1.00)	0.92 (1.00)	1.35 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	4.60 (1.00)	3.01 (1.00)	4.30 (1.00)
	Prob. Median	3.54 (1.00)	3.66 (1.00)	4.04 (1.00)
	$Persist_{RG,18}$	1.82 (1.00)	1.55 (1.00)	2.01 (1.00)
	$Persist_{RG,24}$	1.26 (1.00)	0.94 (1.00)	1.26 (1.00)
	$Const_{0mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00
θ Odds Ratio	Most Likely	20.22 (1.00)	7.89 (1.00)	12.69 (1.00)
	Prob. Median	10.38 (1.00)	8.51 (1.00)	10.33 (1.00)
	$Persist_{RG,18}$	4.29 (1.00)	3.03 (1.00)	5.66 (1.00)
	$Persist_{RG,24}$	1.78 (1.00)	0.87(1.00)	1.71 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.3a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 8.0mm

Ordinary Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
H Hit Rate	Most Likely	0.82 (0.66)	0.82 (0.60)	0.80 (0.62)
	Prob. Median	0.78 (0.67)	0.78 (0.62)	0.79 (0.62)
	$Persist_{RG,18}$	0.73 (0.67)	0.71 (0.58)	0.73 (0.62)
	$Persist_{RG,24}$	0.74 (0.71)	0.60 (0.59)	0.67 (0.64)
	$Const_{0mm}$	0.80 (0.80)	0.71 (0.71)	0.76 (0.76)
	$Const_{5mm}$	0.80 (0.80)	0.71 (0.71)	0.76 (0.76)
	Climatology	0.69	0.59	0.51
CSI Critical Success Index	Most Likely	0.40 (0.12)	0.50 (0.29)	0.45 (0.15)
	Prob. Median	0.31 (0.12)	0.40 (0.31)	0.41 (0.14)
	$Persist_{RG,18}$	0.21 (0.12)	0.35 (0.25)	0.31 (0.15)
	$Persist_{RG,24}$	0.16 (0.10)	0.17 (0.26)	0.18 (0.13)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.11	0.17	0.86
FAR False Alarm Rate	Most Likely	0.47 (0.80)	0.29 (0.71)	0.41 (0.76)
	Prob. Median	0.56 (0.80)	0.33 (0.71)	0.43 (0.76)
	$Persist_{RG,18}$	0.67 (0.80)	0.50 (0.71)	0.55 (0.76)
	$Persist_{RG,24}$	0.69 (0.80)	0.70 (0.71)	0.68 (0.76)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	0.80	0.71	0.76
POD Probability of Detection	Most Likely	0.62 (0.23)	0.62 (0.26)	0.65 (0.52)
	Prob. Median	0.50 (0.22)	0.50 (0.22)	0.60 (0.55)
	$Persist_{RG,18}$	0.38 (0.22)	0.54 (0.32)	0.50 (0.39)
	$Persist_{RG,24}$	0.25 (0.16)	0.29 (0.28)	0.30 (0.43)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.20	0.29	0.24
B Bias Ratio	Most Likely	1.19 (1.19)	0.88 (0.88)	1.10 (1.10)
	Prob. Median	1.12 (1.13)	0.75 (0.75)	1.05 (1.05)
	$Persist_{RG,18}$	1.12 (1.13)	1.08 (1.08)	1.10 (1.10)
	$Persist_{RG,24}$	0.81 (0.81)	0.96 (0.96)	0.95 (0.95)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.3b **Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 8.0mm**

Relative Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
HSS Heidke Skill Score	Most Likely	0.46 (0.00)	0.54 (0.00)	0.49 (0.00)
	Prob. Median	0.33 (0.00)	0.43 (0.00)	0.45 (0.00)
	$Persist_{RG,18}$	0.18 (0.00)	0.31 (0.00)	0.30 (0.00)
	$Persist_{RG,24}$	0.12 (0.00)	0.02 (0.00)	0.09 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.49 (0.00)	0.52 (0.00)	0.50 (0.00)
	Prob. Median	0.35 (0.00)	0.40 (0.00)	0.45 (0.00)
	$Persist_{RG,18}$	0.19 (0.00)	0.32 (0.00)	0.31 (0.00)
	$Persist_{RG,24}$	0.11 (0.00)	0.02 (0.00)	0.09 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.30 (0.00)	0.37 (0.00)	0.32 (0.00)
	Prob. Median	0.20 (0.00)	0.27 (0.00)	0.29 (0.00)
	$Persist_{RG,18}$	0.10 (0.00)	0.18 (0.00)	0.17 (0.00)
	$Persist_{RG,24}$	0.07 (0.00)	0.01 (0.00)	0.05 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	4.58 (1.00)	6.04 (1.00)	4.48 (1.00)
	Prob. Median	3.30 (1.00)	4.83 (1.00)	4.13 (1.00)
	$Persist_{RG,18}$	2.06 (1.00)	2.42 (1.00)	2.58 (1.00)
	$Persist_{RG,24}$	1.83 (1.00)	1.06 (1.00)	1.43 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	2.30 (1.00)	2.39 (1.00)	2.44 (1.00)
	Prob. Median	1.70 (1.00)	1.79 (1.00)	2.14 (1.00)
	$Persist_{RG,18}$	1.31 (1.00)	1.69 (1.00)	1.61 (1.00)
	$Persist_{RG,24}$	1.15 (1.00)	1.02 (1.00)	1.13 (1.00)
	$Const_{0mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	$Const_{5mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00
θ Odds Ratio	Most Likely	10.56 (1.00)	14.44 (1.00)	10.94 (1.00)
	Prob. Median	5.60 (1.00)	8.67 (1.00)	8.83 (1.00)
	$Persist_{RG,18}$	2.70 (1.00)	4.09 (1.00)	4.17 (1.00)
	$Persist_{RG,24}$	2.11 (1.00)	1.08 (1.00)	1.62 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.4a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 12.0mm

Ordinary Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
H Hit Rate	Most Likely	0.87 (0.83)	0.83 (0.78)	0.83 (0.75)
	Prob. Median	0.85 (0.84)	0.82 (0.79)	0.83 (0.75)
	$Persist_{RG,18}$	0.85 (0.79)	0.82 (0.72)	0.87 (0.78)
	$Persist_{RG,24}$	0.87 (0.83)	0.80 (0.77)	0.77 (0.76)
	$Const_{0mm}$	0.89 (0.89)	0.85 (0.85)	0.85 (0.85)
	$Const_{5mm}$	0.89 (0.89)	0.85 (0.85)	0.85 (0.85)
	Climatology	0.80	0.75	0.75
CSI Critical Success Index	Most Likely	0.15 (0.05)	0.18 (0.06)	0.26 (0.08)
	Prob. Median	0.08 (0.04)	0.12 (0.06)	0.26 (0.08)
	$Persist_{RG,18}$	0.25 (0.06)	0.29 (0.09)	0.31 (0.07)
	$Persist_{RG,24}$	0.15 (0.05)	0.16 (0.07)	0.10 (0.08)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.06	0.08	0.08
FAR False Alarm Rate	Most Likely	0.67 (0.89)	0.62 (0.85)	0.58 (0.85)
	Prob. Median	0.80 (0.89)	0.71 (0.85)	0.58 (0.85)
	$Persist_{RG,18}$	0.64 (0.89)	0.60 (0.85)	0.44 (0.85)
	$Persist_{RG,24}$	0.67 (0.89)	0.70 (0.85)	0.82 (0.85)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	0.89	0.85	0.85
POD Probability of Detection	Most Likely	0.22 (0.07)	0.25 (0.10)	0.42 (0.15)
	Prob. Median	0.11 (0.06)	0.17 (0.09)	0.42 (0.15)
	$Persist_{RG,18}$	0.44 (0.13)	0.50 (0.18)	0.42 (0.11)
	$Persist_{RG,24}$	0.22 (0.07)	0.25 (0.12)	0.17 (0.13)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.11	0.15	0.15
B Bias Ratio	Most Likely	0.67 (0.67)	0.67 (0.67)	1.00 (1.00)
	Prob. Median	0.56 (0.56)	0.58 (0.58)	1.00 (1.00)
	$Persist_{RG,18}$	1.22 (1.22)	1.25 (1.25)	0.75 (0.75)
	$Persist_{RG,24}$	0.67 (0.67)	0.83 (0.83)	0.92 (0.92)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.4b **Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Total > 12.0mm**

Relative Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
HSS Heidke Skill Score	Most Likely	0.20 (0.00)	0.21 (0.00)	0.32 (0.00)
	Prob. Median	0.07 (0.00)	0.12 (0.00)	0.32 (0.00)
	$Persist_{RG,18}$	0.32 (0.00)	0.34 (0.00)	0.40 (0.00)
	$Persist_{RG,24}$	0.20 (0.00)	0.16 (0.00)	0.04 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.17 (0.00)	0.18 (0.00)	0.32 (0.00)
	Prob. Median	0.06 (0.00)	0.10 (0.00)	0.32 (0.00)
	$Persist_{RG,18}$	0.35 (0.00)	0.37 (0.00)	0.36 (0.00)
	$Persist_{RG,24}$	0.17 (0.00)	0.15 (0.00)	0.04 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.11 (0.00)	0.12 (0.00)	0.19 (0.00)
	Prob. Median	0.04 (0.00)	0.06 (0.00)	0.19 (0.00)
	$Persist_{RG,18}$	0.19 (0.00)	0.20 (0.00)	0.25 (0.00)
	$Persist_{RG,24}$	0.11 (0.00)	0.09 (0.00)	0.02 (0.00)
	$Const_{0mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{5mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	4.06 (1.00)	3.50 (1.00)	4.17 (1.00)
	Prob. Median	2.03 (1.00)	2.33 (1.00)	4.17 (1.00)
	$Persist_{RG,18}$	4.63 (1.00)	3.89 (1.00)	7.29 (1.00)
	$Persist_{RG,24}$	4.06 (1.00)	2.50 (1.00)	1.30 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	1.22 (1.00)	1.24 (1.00)	1.54 (1.00)
	Prob. Median	1.06 (1.00)	1.11 (1.00)	1.54 (1.00)
	$Persist_{RG,18}$	1.63 (1.00)	1.74 (1.00)	1.62 (1.00)
	$Persist_{RG,24}$	1.22 (1.00)	1.20 (1.00)	1.05 (1.00)
	$Const_{0mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	$Const_{5mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00
θ Odds Ratio	Most Likely	4.93 (1.00)	4.33 (1.00)	6.43 (1.00)
	Prob. Median	2.16 (1.00)	2.60 (1.00)	6.43 (1.00)
	$Persist_{RG,18}$	7.54 (1.00)	6.78 (1.00)	11.79 (1.00)
	$Persist_{RG,24}$	4.93 (1.00)	3.00 (1.00)	1.36 (1.00)
	$Const_{0mm}$	* (*)	* (*)	* (*)
	$Const_{5mm}$	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

5.4.4 Assessment of Probability Forecasts of Accumulations

One of the potentially important parts of the Evening Update forecasts is its probability forecast content. This section outlines an analysis which assesses how well the probability forecasts have performed.

The analysis here uses a performance measure appropriate to probability forecasts and compares the results found for the Evening Update forecasts with certain other forecasts. The performance measure can equally-well be applied to single-valued forecasts, where the forecast is treated as expressing absolute certainty in a single value. In this case the performance measure is directly equivalent to the usual Mean Absolute Error statistic. Table 5.4.1.1 lists the single-valued forecasts used here. This is essentially the same set of forecasts used for the direct analysis of single-valued forecasts in Section 5.4.3. In addition, as outlined in Section 5.4.1 and Table 5.4.1.1, a set of probability forecasts have been created for comparison with those in the Evening Update by taking the single-valued forecasts and attaching a somewhat arbitrary uncertainty-band: when the forecast amount is moderately large, this band extends from 0 up to twice the central forecast amount. The specification of this uncertainty band has not been subjected to detailed consideration and is simply put forward for comparison against the performance of the Evening Update probability forecasts.

The results of the analysis of the probability forecasts are given in Table 5.4.4.1. The upper part of the table relates to the performance of the single-valued forecasts when treated as expressing absolute certainty. Values here are identical to those for the Mean Absolute Error given in Table 5.4.3.1.1 and they are repeated here because the Continuous Brier Score is identical to the Mean Absolute Error when a single-valued forecast is treated as absolutely certain. The lower part of the Table gives the Continuous Brier Score for the constructed probability forecasts and for the Evening Updates' probability forecasts. It can be seen that including the uncertainty band with the single-valued forecasts has always decreased the performance measure in these cases. However, note that adding uncertainty of greater amounts would eventually lead to an increase in the score. The results for the Evening Updates' probability forecasts are somewhat disappointing in comparison with those for the constructed probability forecasts, particularly when considering the probability forecast obtained from the "Most Likely" forecast by adding a simple uncertainty band. It seems that the probability forecasts contained in the Evening Updates are not much better, if at all better, than could be obtained by a simple uncertainty band centred about the main forecast-value.

Tables 5.4.4.2 and 5.4.4.3 relate directly to the question of whether there is enough evidence in the test dataset to distinguish between the performances of the different types of probability forecast. Taking the 'most likely' forecast ("Most Likely") from the Evening Updates, with the addition of either zero or 100% uncertainty, as a "base forecast", Table 5.4.4.2 considers each of the other forecast sources in turn and asks how much evidence there is that the "base forecast" is better than the candidate forecast. In Table 5.4.4.3 the "base forecast" is the probability forecast contained in the Evening Updates. The values in these tables are the standardised differences discussed earlier in Section 2.2.5, and positive values indicate that the "base forecast" has a better performance, as measured by the Continuous Brier Score, than the

candidate. If the candidate forecast had a better performance, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the Continuous Brier Scores for the two forecast sources will turn out to be in the order indicated. For the purposes here, a standardised difference outside the range ± 2 units indicates fairly strong evidence that one forecast source really is better than another.

The results shown in Tables 5.4.4.2 and 5.4.4.3 can be interpreted as follows. Firstly, both the operational probability forecast and the probability forecast derived by adding a 100% uncertainty band to the 'Most Likely' forecast are better than the probability forecasts constructed by attaching 100% uncertainty bands to the persistence forecasts or the constant-valued forecasts. However, the size of the test dataset is too small to allow a clear distinction to be made between the operational probability forecast from the Evening Updates and the simple type of probability forecast derived by adding a 100% uncertainty band to the 'Most Likely' forecast that is contained in the Evening Updates: but the results here favour the latter.

The above conclusion about the probability forecasts in the Evening Updates needs to be tempered by the considerations that the probability forecasts in the Evening Updates are not given to a high resolution and it may be that if a finer resolution had been used, better results might have been obtained.

Table 5.4.4.1 Assessment measures for Probability Forecasts associated with Evening Update forecasts in the Thames Region. (Rainfall Totals)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error (mm)	*(certain)*			
	Most Likely	3.73	3.50	3.83
	Prob. Median	3.88	3.61	3.94
	$Persist_{RG,18}$	5.26	5.27	5.46
	$Persist_{RG,24}$	5.60	6.44	6.24
	$Const_{0mm}$	5.08	5.59	6.00
	$Const_{5mm}$	4.76	5.12	4.99
Continuous Brier Score (mm)	*(100% error)*			
	Most Likely	2.84	2.60	2.87
	Prob. Median	2.88	2.70	3.03
	$Persist_{RG,18}$	3.86	4.00	4.05
	$Persist_{RG,24}$	4.35	4.97	4.60
	$Const_{0mm}$	4.78	5.28	5.67
	$Const_{5mm}$	3.57	3.70	3.73
	(operational)			
	Prob. Forecast	2.90	2.70	3.02

Table 5.4.4.2 Comparison of forecast sources: Standardised Differences for Assessment Measures for Probability Forecasts associated with Evening Update forecasts in the Thames Region. (Rainfall Totals) (In this Table, the base forecast is "Most Likely" with either zero or 100% error)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error (mm)	*(certain)*			
	Prob. Median	0.73	0.35	0.44
	$Persist_{RG,18}$	2.06	2.62	2.48
	$Persist_{RG,24}$	2.83	4.37	3.33
	$Const_{0mm}$	2.67	3.22	4.04
	$Const_{5mm}$	2.31	3.28	2.65
Continuous Brier Score (mm)	*(100% error)*			
	Prob. Median	0.33	0.39	0.88
	$Persist_{RG,18}$	2.17	3.50	3.06
	$Persist_{RG,24}$	3.20	4.39	3.45
	$Const_{0mm}$	4.09	4.30	5.44
	$Const_{5mm}$	2.34	2.56	2.56
	(operational)			
	Prob. Forecast	0.48	0.42	0.95

Table 5.4.4.3 Comparison of forecast sources: Standardised Differences for Assessment Measures for Probability Forecasts associated with Evening Update forecasts in the Thames Region. (Rainfall Totals) (In this Table, the base forecast is "Prob. Forecast": the operational probability forecast)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Continuous Brier Score (mm)	*(100% error)*			
	Most Likely	-0.48	-0.42	-0.95
	Prob. Median	-0.24	0.01	0.04
	$Persist_{RG,18}$	2.05	3.05	2.53
	$Persist_{RG,24}$	3.23	5.52	3.36
	$Const_{0mm}$	4.07	5.02	5.11
	$Const_{5mm}$	2.48	3.59	2.23

5.4.5 Assessment of Single-valued Forecasts of Rates

5.4.5.1 Assessment of forecasts of maximum rates

The analysis here for forecasts of rainfall rates follows the same outline as used in Section 5.4.3.1 for forecasts of rainfall amounts. However, as discussed in Section 5.4.1, fewer forecast-occasions are available for the analysis for this forecast quantity than are available for rainfall amounts. An extra complication for the present case is that there are two potential sources of "ground-truth", deriving either from a network of raingauges or from weather radar. The extra source of ground-truth has prompted the introduction of a further two types of persistence forecast, so that the versions of ground-truth are treated on an equitable basis. It should be recalled that the quantities being forecasted here relate to the maximum rainfall rate experienced at any time in an 18-hour time-period and at any location within a given sub-area of Thames Region of the Environment Agency.

Table 5.4.5.1.1 shows the basic assessment measures for the size of forecast errors for rainfall rates, evaluated for the 3 sub-areas of the Thames region. Results are given for the 8 types for forecasts listed in Table 5.4.1.2. As in Section 5.4.3.1, results are given for the best performance measure obtainable by a constant-value forecast (rows labelled "$Const_{best}$"). Table 5.4.5.1.2 shows the corresponding R^2 (efficiency) measures: these effectively compare the values of the performance measures shown in Table 5.4.5.1.1 with the best performance measure achievable by a constant-value forecast: that is, they compare the performance measures, as given in Table 5.4.5.1.1, for the given forecast source with the corresponding results for "$Const_{best}$" .

The results in Table 5.4.5.1.1 immediately indicate that the forecasts contained within the Evening Updates are considerably better matched to the ground-truth obtained from the raingauge network than they are to that from the weather radar source used here. Section 5.4.1 has outlined the potential problems with data from this radar source and results here are to be treated with caution. Examination of the example data in Table 5.4.2.2 shows that the spatial maxima obtained from the radar source are typically much larger than those found from the raingauge network. Some of the difference in forecast performance between these sources that is shown in Table 5.4.5.1.1 arises from this fact. The persistence forecasts obtained from the radar source have, comparatively, a very poor performance in forecasting the raingauge-derived ground-truth, and their performance in forecasting the radar-derived ground-truth is in line with the other forecast sources. This indicates that the difference in forecast performance for the two ground-truths is not fully explained simply by the size of the target quantities.

The results for the raingauge-derived ground-truth in Tables 5.4.5.1.1 and 5.4.5.1.2 are broadly similar to those found for forecasts of rainfall amounts, except for the size of the errors and slightly reduced R^2-values. Similarly to the finding for forecasts of rainfall amounts, the performance measures for the different sources of forecasts of rainfall rates have the expected ranking, with the forecasts from the Evening Updates being better than both persistence forecasts and constant-value forecasts. As might be expected, 18-hour-delayed persistence forecasts are usually better than the 24-hour-delayed persistence forecasts. However, the R^2 (efficiency) measures for the

persistence forecasts are usually negative, indicating that a better forecast performance than provided by the persistence forecasts can be achieved by selecting a suitable constant value to use as a constant-value forecast. Again, there is little difference in performance between the two forecasts obtained from the Evening Updates. It seems that the values taken directly from the forecasts, given by the 'most likely' values, are slightly better than the forecast derived as the median of the probability forecasts when ordinary errors are used but slightly worse for performance measures based on errors of logarithms.

Table 5.4.5.1.3 shows details of the bias contained in the various forecast sources and compared with the two-versions of ground-truth. This clearly reveals the difference in the biases associated with the ground-truths. Overall it seems that the Evening Update forecasts give values which are slightly too small (compared with the raingauge-derived ground truth), with the forecasts derived as the median of the probability forecasts tending to be somewhat smaller than the 'most likely' values. Table 5.4.5.1.4 shows some statistics for the rainfall amounts which give more details of the typical amounts obtained for the actual outcomes and for the forecasts. In contrast to the results found for Table 5.4.3.1.4, this table shows that the variation between the sub-areas of the typical amounts observed for the outcome is not followed by the variation in the typical amounts being forecasted in the Evening Updates: this may be simply a sample-size related effect. The results here show that the outcome values derived from radar have much higher standard-deviations than those derived from raingauges, as well as much larger means and medians. As far as can be traced, the apparent factor of four between the two sources of ground-truth does not seem to be related to a problem in converting data between different measurement units, as might be suspected, but is directly associated with the finer spatial resolution of the radar data and with the high spatial variability of the radar fields derived from weather radar.

Table 5.4.5.1.5 gives values for correlation and regression coefficients for linear relationships between outcomes and forecasts of rainfall and log-rainfall. The values here indicate that modest improvements may be obtained to the raw forecasts already considered by forming a simple adjustment of the form:

$$f_i^* = \mu_o + \beta(f_i - \mu_f).$$

Here, f_i^* is the new forecast value constructed from the raw value f_i for occasion i, β is the regression coefficient and μ_o and μ_f are the means of the outcome and raw forecast values. The extent of potential improvement can be judged by comparing the square of the correlation coefficient with the value, in Table 5.4.5.1.2, for R^2 for Root Mean Square Error. For example, the R^2 for the "Most Likely" forecast of radar groud-truth in the NE sub-area might be increased from -0.19 to 0.14 for a simple scaling of the rainfall amount, or, in the log-space, from -0.24 to 0.35 for an adjustment to the logarithm of rainfall amounts. Here, the formula for the new forecast would be

$$f_i^* = 32.1 + 2.23(f_i - 6.75)$$

when using rainfall directly. The smallest value forecasted would be 17mm h^{-1}. The impression given by the earlier tables may have been that the ground-truth derived from radar cannot be forecasted because the values are mainly noise reflecting random short-term fluctuations in the radar-images. Table 5.4.5.1.5 partly contradicts

this by revealing that the correlations of the Evening Update forecast with the radar-derived ground-truth are moderately high and only slightly lower than those of the forecasts with the raingauge-derived ground-truth.

Table 5.4.5.1.6 relates to the question of whether the evidence provided by the test dataset is sufficient to distinguish between the performance of different forecast sources, bearing in mind the sampling variability of the forecast performance statistics and the statistical dependences between them. Taking the 'most likely' forecast ("Most Likely") from the Evening Updates as a "base forecast", Table 5.4.5.1.6 considers each of the other forecast sources in turn and asks how much evidence there is that the "base forecast" is better than the candidate forecast. Once again, the values given are the standardised differences discussed earlier in Section 4.3.3, and positive values indicate that the "base forecast" has a smaller size of error, as measured by the performance statistic, than the candidate. If the candidate forecast produces smaller errors, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the long-run performance measures for the two forecast sources will turn out to be in the order indicated. For the purposes here a standardised difference outside the range ± 2 units indicates fairly strong evidence that one forecast source really is better than another.

The results in Table 5.4.5.1.6 reflect those in Table 5.4.5.1.1, in that the comparisons which favour one forecast source over another are the same. However, Table 5.4.5.1.6 provides extra information. For example, it shows that there is only weak evidence that the 'most likely' values in the Evening Update provide better or worse forecasts of the raingauge-derived ground-truth than the median-values derived from the probability forecasts, and that the apparent preference differs between the measures of forecast performance. There is only weak evidence that the "Most Likely" forecast is better than a persistence forecast based on the 18-hours immediately before the start of the forecast period. (It should be recalled that the Evening Updates are typically issued 2 hours before the start of the forecast period.)

Once again, it appears from Table 5.4.5.1.6 that the performance measures based on errors in the logarithms of rainfall amounts are able to provide stronger evidence in the comparison of forecast sources than do those based on ordinary errors. Similarly, the Root Mean Square Error performance measure appears to provide weaker evidence for differences than does the Mean Absolute Error. These are the same as the findings in Section 5.4.3 for forecasts of rainfall amounts (see the discussion there).

Table 5.4.5.1.1 Raw assessment measures for Evening Update forecasts in the Thames Region. (Rainfall Rates)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error (mm/hr)	Most Likely	6.84	7.68	5.28	26.3	24.7	25.7
	Prob. Median	7.03	7.41	5.61	26.9	25.0	26.2
	$Persist_{RG,18}$	9.45	8.77	6.55	24.4	23.4	26.3
	$Persist_{RG,24}$	9.31	11.2	7.47	27.6	27.4	26.3
	$Persist_{RD,18}$	28.5	29.6	25.2	28.5	26.3	22.7
	$Persist_{RD,24}$	28.9	26.4	27.3	37.1	26.2	28.8
	$Const_{0mm/hr}$	8.85	8.78	8.38	32.1	30.6	32.6
	$Const_{10mm/hr}$	9.28	9.02	8.17	27.8	23.8	26.7
	$Const_{best}$	7.82	7.82	6.97	27.5	22.6	25.8
Root Mean Square Error (mm/hr)	Most Likely	11.5	12.8	9.00	51.5	41.3	47.6
	Prob. Median	12.2	13.0	9.62	52.1	41.9	48.0
	$Persist_{RG,18}$	15.1	14.2	11.1	47.8	38.1	47.9
	$Persist_{RG,24}$	13.8	17.5	12.0	51.5	42.9	46.4
	$Persist_{RD,18}$	48.3	47.2	46.7	47.5	37.5	44.8
	$Persist_{RD,24}$	50.4	40.4	47.3	59.9	38.4	47.1
	$Const_{0mm/hr}$	14.9	14.7	13.3	57.0	46.6	53.8
	$Const_{10mm/hr}$	12.1	11.9	10.5	52.0	40.7	48.4
	$Const_{best}$	12.0	11.9	10.4	47.1	35.2	42.8
Mean Absolute Error of Log-Rainfall (dim'less)	Most Likely	0.97	1.08	0.70	1.45	1.61	1.47
	Prob. Median	0.95	0.99	0.76	1.56	1.78	1.57
	$Persist_{RG,18}$	1.05	1.00	0.90	1.43	1.89	1.81
	$Persist_{RG,24}$	1.32	1.53	1.20	1.78	2.07	1.79
	$Persist_{RD,18}$	1.65	1.82	1.55	1.32	1.25	1.07
	$Persist_{RD,24}$	1.81	1.90	1.73	1.62	1.20	1.25
	$Const_{0mm/hr}$	2.15	2.11	2.22	3.30	3.67	3.58
	$Const_{10mm/hr}$	1.51	1.54	1.37	1.40	1.12	1.17
	$Const_{best}$	1.31	1.37	1.21	1.37	1.01	1.10
Root Mean Square Error of Log-Rainfall (dim'less)	Most Likely	1.20	1.46	0.92	1.86	2.06	1.84
	Prob. Median	1.19	1.39	0.99	1.97	2.23	1.93
	$Persist_{RG,18}$	1.36	1.40	1.19	1.88	2.28	2.22
	$Persist_{RG,24}$	1.65	2.03	1.62	2.21	2.53	2.22
	$Persist_{RD,18}$	1.92	2.15	1.82	1.75	1.53	1.49
	$Persist_{RD,24}$	2.16	2.33	2.11	2.02	1.49	1.65
	$Const_{0mm/hr}$	2.61	2.62	2.62	3.70	3.90	3.85
	$Const_{10mm/hr}$	1.82	1.91	1.72	1.67	1.40	1.45
	$Const_{best}$	1.48	1.56	1.40	1.67	1.32	1.41

Table 5.4.5.1.2 R^2 (efficiency) measures for Evening Update forecasts in the Thames Region for each type of assessment measure. (Rainfall Rates)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
R^2 for Mean Absolute Error	Most Likely	0.12	0.02	0.24	0.04	-0.09	0.00
	Prob. Median	0.10	0.05	0.20	0.02	-0.11	-0.01
	$Persist_{RG,18}$	-0.21	-0.12	0.06	0.11	-0.04	-0.02
	$Persist_{RG,24}$	-0.19	-0.43	-0.07	-0.01	-0.21	-0.02
	$Persist_{RD,18}$	-2.65	-2.79	-2.61	-0.04	-0.17	0.12
	$Persist_{RD,24}$	-2.70	-2.37	-2.91	-0.35	-0.16	-0.11
	$Const_{0mm/hr}$	-0.13	-0.12	-0.20	-0.17	-0.35	-0.26
	$Const_{10mm/hr}$	-0.19	-0.15	-0.17	-0.01	-0.06	-0.03
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00
R^2 for Root Mean Square Error	Most Likely	0.08	-0.17	0.25	-0.19	-0.38	-0.24
	Prob. Median	-0.03	-0.21	0.14	-0.22	-0.42	-0.26
	$Persist_{RG,18}$	-0.57	-0.44	-0.15	-0.03	-0.17	-0.25
	$Persist_{RG,24}$	-0.32	-1.19	-0.34	-0.19	-0.49	-0.17
	$Persist_{RD,18}$	-15.2	-14.9	-19.2	-0.02	-0.14	-0.10
	$Persist_{RD,24}$	-16.6	-10.7	-19.8	-0.62	-0.19	-0.21
	$Const_{0mm/hr}$	-0.54	-0.55	-0.65	-0.46	-0.76	-0.58
	$Const_{10mm/hr}$	-0.01	-0.01	-0.02	-0.22	-0.34	-0.28
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00
R^2 for Mean Absolute Error of Log-Rainfall	Most Likely	0.26	0.22	0.42	-0.06	-0.59	-0.34
	Prob. Median	0.27	0.28	0.37	-0.14	-0.75	-0.43
	$Persist_{RG,18}$	0.20	0.27	0.26	-0.04	-0.86	-0.65
	$Persist_{RG,24}$	-0.01	-0.11	0.01	-0.30	-1.04	-0.63
	$Persist_{RD,18}$	-0.26	-0.32	-0.28	0.03	-0.23	0.02
	$Persist_{RD,24}$	-0.38	-0.38	-0.43	-0.19	-0.19	-0.14
	$Const_{0mm/hr}$	-0.65	-0.53	-0.83	-1.42	-2.62	-2.26
	$Const_{10mm/hr}$	-0.15	-0.12	-0.13	-0.02	-0.10	-0.07
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00
R^2 for Root Mean Square Error of Log-Rainfall	Most Likely	0.34	0.11	0.57	-0.24	-1.44	-0.71
	Prob. Median	0.35	0.21	0.50	-0.39	-1.84	-0.89
	$Persist_{RG,18}$	0.15	0.18	0.28	-0.26	-1.97	-1.50
	$Persist_{RG,24}$	-0.25	-0.70	-0.33	-0.76	-2.67	-1.49
	$Persist_{RD,18}$	-0.69	-0.91	-0.69	-0.10	-0.34	-0.12
	$Persist_{RD,24}$	-1.15	-1.24	-1.27	-0.46	-0.27	-0.38
	$Const_{0mm/hr}$	-2.12	-1.84	-2.50	-3.91	-7.71	-6.50
	$Const_{10mm/hr}$	-0.52	-0.51	-0.51	0.00	-0.12	-0.07
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.4.5.1.3 Bias measures for Evening Update forecasts in the Thames Region . (Rainfall Rates)

Bias Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Error (mm/hr)	Most Likely	2.10	1.86	1.41	25.3	23.6	25.6
	Prob. Median	2.69	2.54	1.94	25.9	24.3	26.2
	$Persist_{RG,18}$	-4.35	-1.43	1.21	18.9	20.3	25.4
	$Persist_{RG,24}$	0.02	-0.02	0.14	23.3	21.7	24.4
	$Persist_{RD,18}$	-26.5	-25.6	-23.9	-3.29	-3.88	0.32
	$Persist_{RD,24}$	-23.4	-22.6	-24.3	-0.12	-0.81	-0.08
	$Const_{0mm/hr}$	8.85	8.78	8.38	32.1	30.5	32.6
	$Const_{10mm/hr}$	-1.15	-1.22	-1.62	22.1	20.5	22.6
Median Error (mm/hr)	Most Likely	0.60	0.00	0.80	5.81	10.3	8.25
	Prob. Median	0.80	0.00	0.80	7.53	12.7	10.6
	$Persist_{RG,18}$	-0.80	0.00	0.00	2.49	11.2	10.5
	$Persist_{RG,24}$	0.00	0.00	0.00	9.63	12.2	10.1
	$Persist_{RD,18}$	-9.30	-10.4	-8.45	-0.59	-1.78	0.41
	$Persist_{RD,24}$	-8.48	-12.9	-11.0	0.28	-0.62	2.25
	$Const_{0mm/hr}$	3.20	3.20	4.00	16.03	16.03	14.66
	$Const_{10mm/hr}$	-6.80	-6.80	-6.00	6.03	6.03	4.66
Mean Error of Log-Rainfall (dim'less)	Most Likely	0.06	-0.01	0.10	1.22	1.55	1.46
	Prob. Median	0.28	0.17	0.20	1.44	1.73	1.57
	$Persist_{RG,18}$	-0.30	0.05	0.26	0.86	1.61	1.62
	$Persist_{RG,24}$	0.01	0.01	0.08	1.17	1.57	1.44
	$Persist_{RD,18}$	-1.23	-1.42	-1.27	-0.07	0.14	0.10
	$Persist_{RD,24}$	-1.20	-1.61	-1.37	-0.05	-0.05	-0.01
	$Const_{0mm/hr}$	2.15	2.11	2.22	3.30	3.67	3.58
	$Const_{10mm/hr}$	-1.07	-1.11	-1.00	0.08	0.45	0.36
Median Error of Log-Rainfall (dim'less)	Most Likely	0.18	0.00	0.27	1.16	1.43	1.14
	Prob. Median	0.47	0.00	0.41	1.56	1.59	1.25
	$Persist_{RG,18}$	-0.25	0.00	0.00	0.96	1.75	1.54
	$Persist_{RG,24}$	0.00	0.00	0.00	1.17	1.83	1.45
	$Persist_{RD,18}$	-1.36	-1.49	-1.42	-0.10	-0.27	0.03
	$Persist_{RD,24}$	-1.47	-1.51	-1.23	-0.11	-0.03	0.15
	$Const_{0mm/hr}$	2.08	2.08	2.30	3.69	3.69	3.60
	$Const_{10mm/hr}$	-1.14	-1.14	-0.92	0.47	0.47	0.38

Table 5.4.5.1.4 Statistics of forecasts and outcomes for Evening Update forecasts in Thames Region. (Rainfall Rates)

Statistic of Rainfall	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Rainfall (mm/hr)	Outcome (Raingauge)	8.85	8.78	8.38
	Outcome (Radar)	32.1	30.5	32.6
	Most Likely	6.75	6.94	6.97
	Prob. Median	6.16	6.24	6.44
Median Rainfall (mm/hr)	Outcome (Raingauge)	3.20	3.20	4.00
	Outcome (Radar)	16.0	16.0	14.7
	Most Likely	4.00	4.00	4.00
	Prob. Median	5.00	4.00	5.00
Standard Deviation (mm/hr)	Outcome (Raingauge)	12.1	11.9	10.5
	Outcome (Radar)	47.6	35.5	43.2
	Most Likely	7.90	8.14	8.05
	Prob. Median	7.96	7.89	7.58

Table 5.4.5.1.5 Correlation of Evening Update forecasts with outcomes in Thames Region. (Rainfall Rates)

Correlation Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Correlation	Most Likely	0.41	0.23	0.56	0.37	0.26	0.42
	Prob. Median	0.34	0.20	0.48	0.33	0.24	0.41
	$Persist_{RG,18}$	0.58	0.51	0.50	0.36	0.40	0.33
	$Persist_{RG,24}$	0.34	-0.09	0.34	0.23	0.01	0.43
	$Persist_{RD,18}$	0.50	0.42	0.60	0.47	0.56	0.48
	$Persist_{RD,24}$	0.32	0.29	0.32	0.19	0.40	0.40
Regression Coefficient	Most Likely	0.63	0.34	0.72	2.23	1.15	2.26
	Prob. Median	0.52	0.31	0.67	1.95	1.10	2.36
	$Persist_{RG,18}$	0.40	0.38	0.45	0.96	0.90	1.21
	$Persist_{RG,24}$	0.34	-0.09	0.33	0.88	0.03	1.75
	$Persist_{RD,18}$	0.13	0.11	0.14	0.49	0.46	0.46
	$Persist_{RD,24}$	0.08	0.10	0.08	0.19	0.40	0.40
Correlation of Log-Rainfall	Most Likely	0.65	0.50	0.78	0.59	0.48	0.67
	Prob. Median	0.70	0.58	0.76	0.64	0.49	0.68
	$Persist_{RG,18}$	0.63	0.61	0.65	0.48	0.42	0.41
	$Persist_{RG,24}$	0.38	0.17	0.37	0.29	0.07	0.31
	$Persist_{RD,18}$	0.60	0.48	0.58	0.49	0.46	0.46
	$Persist_{RD,24}$	0.32	0.32	0.34	0.24	0.36	0.31
Regression Coefficient of Log-Rainfall	Most Likely	0.70	0.57	0.80	0.72	0.47	0.69
	Prob. Median	0.69	0.63	0.75	0.72	0.45	0.67
	$Persist_{RG,18}$	0.58	0.58	0.66	0.50	0.34	0.41
	$Persist_{RG,24}$	0.57	0.16	0.35	0.33	0.06	0.30
	$Persist_{RD,18}$	0.50	0.47	0.56	0.46	0.38	0.45
	$Persist_{RD,24}$	0.30	0.38	0.34	0.25	0.37	0.31

Table 5.4.5.1.6 Comparison of forecast sources: Standardised differences for assessment measures for Evening Update forecasts in the Thames Region. (Rainfall Rates) (In this Table, the base forecast is "Most Likely")

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error	Prob. Median	0.56	-0.68	0.73	1.65	0.81	0.98
	$Persist_{RG,18}$	1.56	1.19	1.27	-0.96	-0.69	0.51
	$Persist_{RG,24}$	1.89	2.45	1.88	1.02	2.02	0.43
	$Persist_{RD,18}$	4.26	4.84	3.89	0.38	0.56	-0.61
	$Persist_{RD,24}$	4.09	4.58	4.33	2.36	0.39	0.55
	$Const_{0mm/hr}$	1.94	1.32	2.97	5.87	6.07	6.26
	$Const_{10mm/hr}$	2.50	1.31	2.98	1.64	-0.82	0.91
Root Mean Square Error	Prob. Median	1.29	0.51	0.81	1.45	1.14	0.70
	$Persist_{RG,18}$	1.14	1.46	1.37	-1.35	-1.15	0.18
	$Persist_{RG,24}$	1.18	2.07	1.53	0.00	0.95	-0.44
	$Persist_{RD,18}$	2.60	3.24	2.48	-0.44	-0.75	-0.33
	$Persist_{RD,24}$	2.76	2.99	2.67	1.23	-0.49	-0.05
	$Const_{0mm/hr}$	2.05	1.76	2.40	3.27	3.82	3.74
	$Const_{10mm/hr}$	0.45	-0.79	1.26	0.59	-0.54	0.70
Mean Absolute Error of Log-Rainfall	Prob. Median	-0.32	-1.24	1.05	1.61	2.05	1.32
	$Persist_{RG,18}$	0.55	-0.45	1.66	-0.16	1.61	2.49
	$Persist_{RG,24}$	2.08	2.38	3.32	1.94	2.40	1.88
	$Persist_{RD,18}$	4.47	3.92	5.43	-0.69	-1.79	-1.96
	$Persist_{RD,24}$	4.53	3.66	5.76	0.86	-1.81	-0.96
	$Const_{0mm/hr}$	5.18	4.40	7.34	8.92	11.5	11.18
	$Const_{10mm/hr}$	3.27	2.28	4.06	-0.28	-2.55	-1.68
Root Mean Square Error of Log-Rainfall	Prob. Median	-0.08	-1.39	0.99	1.44	1.77	1.16
	$Persist_{RG,18}$	0.99	-0.33	1.92	0.13	1.23	2.68
	$Persist_{RG,24}$	2.27	2.49	3.40	1.78	2.17	1.89
	$Persist_{RD,18}$	3.95	2.99	4.49	-0.62	-2.27	-1.61
	$Persist_{RD,24}$	4.44	3.24	4.53	0.77	-2.10	-0.79
	$Const_{0mm/hr}$	5.46	4.99	6.73	7.81	9.76	9.24
	$Const_{10mm/hr}$	3.66	2.03	4.22	-0.83	-2.70	-1.90

5.4.5.2 Assessment of category-forecasts

The analysis here for forecasts of whether rainfall rates will exceed given thresholds follows the same outline as that used in Section 5.4.3.2 for forecasts of rainfall amounts. Once again (see Section 5.4.5.1) there are fewer forecast-occasions available for analysis of rainfall rates than for rainfall amounts. There are two potential sources of "ground-truth", deriving either from a network of raingauges or from weather radar. For conciseness, for the present set of analyses, the constant-valued forecasts listed in Table 5.4.1.3 have been omitted, leaving 6 candidate forecasts for comparison. As in Section 5.4.3.2, the tables of results include values for the performance measures that would be achieved by two types of random forecast, one based on the observed rate of threshold-exceedence among the outcomes for the given ground-truth and one based on the rate found for the given forecast source.

For the present study a number of thresholds for rainfall rates have been chosen which are perhaps unrealistic for practical use, but they illustrate the problems involved in attempting to specify performance measures for categorical forecasts in circumstances where the numbers of cases is limited.

Tables 5.4.5.2.1 to 5.4.5.2.4 show results for a collection of performance measures for analyses using thresholds of 0, 4, 12 and 25mm h^{-1} for the maximum rainfall rate in the 18 hour forecast period in each of the 3 sub-areas of Thames Region. The types of performance measures available for categorical forecasts fall naturally into two groups, and each table is divided in two corresponding parts. In the first group are the ordinary score statistics in which the performance measures are defined fairly directly in terms of the rates of occurrences of success or failure of the forecasts: these are listed in part **a** of each Table. The second group includes more refined measures in which the forecast performance is measured relative to what could be achieved by random forecasts of the two types outlined above: these are listed in part **b** of each Table.

In constructing the tables of results for performance measures of categorical forecasts, there are many cases where the values cannot be calculated because of the need to divide by zero: in these cases the results are represented by an asterisk (*). This rule has been applied even in cases where the standard formula formally gives 0/0 and where there is a potential to create a meaningful numerical value by re-expressing the formula in an alternative way.

The performance scores for a threshold of 0 mm h^{-1} (Table 5.4.5.2.1) illustrate the problem of defining performance measures in extreme cases: here the given threshold is always exceeded for the observed outcomes of rainfall rate derived from the radar source. This leads to the False Alarm Rate being zero for all forecast sources and to the Hit Rate and Critical Success Index being equal to one for the persistence-based forecasts derived from the same radar source (except for the $Persist_{RD,18}$ forecast for the NE sub-area, for which one of the forecast rates is zero: see Table 5.4.2.2). The zero-counts, arising from the target threshold always being exceeded among the observed outcomes, lead to the relative skill scores being evaluated either as zero or as undefined (denoted by an asterisk). When the network of raingauges is used to provide the ground-truth, the performance statistics appear better behaved. In this case the results for rainfall rates are similar to those for rainfall amounts (see Section

5.4.3.2). Thus, the ordinary scores suggest that the Evening Update forecasts are not substantially better than random forecasts: this impression is contradicted by the relative score measures (Table 5.4.5.2.1b) which suggest that even the persistence forecasts (derived from raingauges) provide a worthwhile improvement over random forecasts. The persistence forecasts derived from the radar source are judged to be poor by all the performance measures when the raingauge ground-truth is used. Once again, the performance measures for a zero-rainfall threshold do not provide any clear ordering among the Evening Update and (raingauge-)persistence forecasts. For example, for the North East sub-area, the persistence forecast taken from the immediately preceding 18 hours is preferred over the "Most Likely" Evening Update forecast according to the Heidke Skill Score, the Equitable Threat Score, the Likelihood Ratio criterion for values not exceeding the threshold and the Odds Ratio. The reverse is true for the Kuipers Skill Score and the Likelihood Ratio criterion for values which do exceed the threshold. Overall, the apparent preference for the various forecasts varies between the 3 sub-areas and between the performance measures being used.

When similar analyses are made for the cases of thresholds at 4 mm h^{-1} and 8 mm h^{-1} (Tables 5.4.5.2.2 and 5.4.5.2.3), the overall conclusions are again unclear. In these cases, most of the performance measures do not reveal any clear distinction between the abilities of the forecasts when matched against either of the two versions of ground truth, in that the 'best' values of the performance measures across the three sub-areas are of a similar size. However, the performance measures vary more widely across sub-areas in the case of the radar-based ground-truth. The performance measures which do distinguish between the ground-truths are the False Alarm Rate and the Bias. Although results vary according to the performance measure and the sub-area selected, there appears to be a slight preference overall for the median derived from the probability forecast in the Evening Updates over the 'Most Likely' value from the same source. It is not clear whether this is simply due to sampling variability. Both of these forecast sources seem to clearly better than the persistence forecasts.

When the threshold on the maximum rainfall rate is raised to 25 mm h^{-1} (Table 5.4.5.2.5), the forecasts from the Evening Update are incorrect on most of the very few (1, 2 or 3) occasions when the threshold is exceeded by the forecast values, at least when judged against the ground-truth from the raingauge network. This leads to very poor values for the performance measures. Note that, for the West sub-area, the 'Most Likely' forecast exceeds the threshold twice and it is correct on one out of the two occasions, whereas the NE and SE sub-areas each have zero out of two correct. This small difference between the sub-areas, largely attributable to sampling variations, is associated with differences in performance measures which can be quite extreme. For example the Odds Ratio is zero for the NE and SE sub-areas and 9.2 for the West sub-area. This highlights the need for an improved procedure to take account of the sampling variability inherent in the performance measures derived from categorical analyses.

Table 5.4.5.2.1a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate > 0.0mm h^{-1}

Ordinary Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
H Hit Rate	Most Likely	0.83 (0.73)	0.81 (0.69)	0.91 (0.75)	0.81 (0.81)	0.83 (0.81)	0.81 (0.81)
	Prob. Median	0.74 (0.63)	0.77 (0.63)	0.83 (0.69)	0.68 (0.68)	0.72 (0.72)	0.74 (0.74)
	$Persist_{RG,18}$	0.91 (0.81)	0.89 (0.69)	0.89 (0.83)	0.92 (0.92)	0.83 (0.83)	0.91 (0.91)
	$Persist_{RG,24}$	0.85 (0.77)	0.81 (0.65)	0.91 (0.78)	0.87 (0.87)	0.75 (0.75)	0.85 (0.85)
	$Persist_{RD,18}$	0.85 (0.85)	0.79 (0.79)	0.91 (0.91)	0.98 (0.98)	1.00 (1.00)	1.00 (1.00)
	$Persist_{RD,24}$	0.87 (0.87)	0.79 (0.79)	0.91 (0.91)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	0.77	0.67	0.83	1.00	1.00	1.00
CSI Critical Success Index	Most Likely	0.82 (0.73)	0.79 (0.68)	0.90 (0.75)	0.81 (0.81)	0.83 (0.81)	0.81 (0.81)
	Prob. Median	0.71 (0.63)	0.74 (0.60)	0.81 (0.68)	0.68 (0.68)	0.72 (0.72)	0.74 (0.74)
	$Persist_{RG,18}$	0.90 (0.81)	0.87 (0.68)	0.88 (0.83)	0.92 (0.92)	0.83 (0.83)	0.91 (0.91)
	$Persist_{RG,24}$	0.84 (0.77)	0.78 (0.63)	0.90 (0.78)	0.87 (0.87)	0.75 (0.75)	0.85 (0.85)
	$Persist_{RD,18}$	0.85 (0.85)	0.79 (0.79)	0.91 (0.91)	0.98 (0.98)	1.00 (1.00)	1.00 (1.00)
	$Persist_{RD,24}$	0.87 (0.87)	0.79 (0.79)	0.91 (0.91)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	0.77	0.66	0.66	1.00	1.00	1.00
FAR False Alarm Rate	Most Likely	0.07 (0.13)	0.14 (0.21)	0.00 (0.09)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Prob. Median	0.06 (0.13)	0.11 (0.21)	0.00 (0.09)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RG,18}$	0.08 (0.13)	0.09 (0.21)	0.06 (0.09)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RG,24}$	0.09 (0.13)	0.10 (0.21)	0.02 (0.09)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RD,18}$	0.13 (0.13)	0.21 (0.21)	0.09 (0.09)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RD,24}$	0.13 (0.13)	0.21 (0.21)	0.09 (0.09)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.13	0.21	0.09	0.00	0.00	0.00
POD Probability of Detection	Most Likely	0.87 (0.81)	0.90 (0.81)	0.90 (0.81)	0.81 (0.81)	0.83 (0.81)	0.81 (0.81)
	Prob. Median	0.74 (0.68)	0.81 (0.68)	0.81 (0.74)	0.68 (0.68)	0.72 (0.72)	0.74 (0.74)
	$Persist_{RG,18}$	0.98 (0.92)	0.95 (0.92)	0.94 (0.91)	0.92 (0.92)	0.83 (0.83)	0.91 (0.91)
	$Persist_{RG,24}$	0.91 (0.87)	0.86 (0.87)	0.92 (0.85)	0.87 (0.87)	0.75 (0.75)	0.85 (0.85)
	$Persist_{RD,18}$	0.98 (0.98)	1.00 (1.00)	1.00 (1.00)	0.98 (0.98)	1.00 (1.00)	1.00 (1.00)
	$Persist_{RD,24}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	0.87	0.79	0.91	1.00	1.00	1.00
B Bias Ratio	Most Likely	0.93 (0.93)	1.05 (1.05)	0.90 (0.90)	0.81 (0.81)	0.83 (0.81)	0.81 (0.81)
	Prob. Median	0.78 (0.78)	0.90 (0.90)	0.81 (0.81)	0.68 (0.68)	0.72 (0.72)	0.74 (0.74)
	$Persist_{RG,18}$	1.07 (1.07)	1.05 (1.05)	1.00 (1.00)	0.92 (0.92)	0.83 (0.83)	0.91 (0.91)
	$Persist_{RG,24}$	1.00 (1.00)	0.95 (0.95)	0.94 (0.94)	0.87 (0.87)	0.75 (0.75)	0.85 (0.85)
	$Persist_{RD,18}$	1.13 (1.13)	1.26 (1.26)	1.10 (1.10)	0.98 (0.98)	1.00 (1.00)	1.00 (1.00)
	$Persist_{RD,24}$	1.15 (1.15)	1.26 (1.26)	1.10 (1.10)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.5.2.1b Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate > 0.0mm h^{-1} Relative Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
HSS Heidke Skill Score	Most Likely	0.37 (0.00)	0.39 (0.00)	0.62 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Prob. Median	0.28 (0.00)	0.39 (0.00)	0.45 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RG,18}$	0.50 (0.00)	0.63 (0.00)	0.34 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RG,24}$	0.34 (0.00)	0.46 (0.00)	0.56 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RD,18}$	-0.03(0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	* (*)	* (*)
	$Persist_{RD,24}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	* (*)	* (*)	* (*)
	Climatology	0.00	0.00	0.00	0.00	*	*
KSS Kuipers Skill Score	Most Likely	0.44 (0.00)	0.36 (0.00)	0.90 (0.00)	* (*)	* (*)	* (*)
	Prob. Median	0.45 (0.00)	0.45 (0.00)	0.81 (0.00)	* (*)	* (*)	* (*)
	$Persist_{RG,18}$	0.41 (0.00)	0.59 (0.00)	0.34 (0.00)	* (*)	* (*)	* (*)
	$Persist_{RG,24}$	0.34 (0.00)	0.49 (0.00)	0.72 (0.00)	* (*)	* (*)	* (*)
	$Persist_{RD,18}$	-0.02(0.00)	0.00(0.00)	0.00 (0.00)	* (*)	* (*)	* (*)
	$Persist_{RD,24}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	* (*)	* (*)	* (*)
	Climatology	0.00	0.00	0.00	*	*	*
ETS Equitable Skill Score	Most Likely	0.23 (0.00)	0.24 (0.00)	0.45 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Prob. Median	0.16 (0.00)	0.24 (0.00)	0.29 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RG,18}$	0.33 (0.00)	0.46 (0.00)	0.20 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RG,24}$	0.21 (0.00)	0.30 (0.00)	0.39 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	$Persist_{RD,18}$	-0.02(0.00)	0.00(0.00)	0.00 (0.00)	0.00 (0.00)	* (*)	* (*)
	$Persist_{RD,24}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	* (*)	* (*)	* (*)
	Climatology	0.00	0.00	0.00	*	*	*
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	2.03 (1.00)	1.66 (1.00)	* (1.00)	* (*)	* (*)	* (*)
	Prob. Median	2.59 (1.00)	2.23 (1.00)	* (1.00)	* (*)	* (*)	* (*)
	$Persist_{RG,18}$	1.71 (1.00)	2.62 (1.00)	1.56 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RG,24}$	1.60 (1.00)	2.36 (1.00)	4.58 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RD,18}$	0.98(1.00)	1.00 (1.00)	1.00 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RD,24}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00	*	*	*
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	4.38 (1.00)	4.77 (1.00)	9.60 (1.00)	* (*)	* (*)	* (*)
	Prob. Median	2.74 (1.00)	3.34 (1.00)	5.33 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RG,18}$	19.7 (1.00)	13.4 (1.00)	6.40 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RG,24}$	4.93 (1.00)	4.45 (1.00)	9.60 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RD,18}$	0.00(1.00)	*(*)	*(*)	* (*)	* (*)	* (*)
	$Persist_{RD,24}$	* (*)	* (*)	* (*)	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00	*	*	*
θ Odds Ratio	Most Likely	8.89 (1.00)	7.92 (1.00)	* (1.00)	* (*)	* (*)	* (*)
	Prob. Median	7.08 (1.00)	7.44 (1.00)	* (1.00)	* (*)	* (*)	* (*)
	$Persist_{RG,18}$	33.7 (1.00)	35.0 (1.00)	10.0 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RG,24}$	7.88 (1.00)	10.5 (1.00)	44.0 (1.00)	* (*)	* (*)	* (*)
	$Persist_{RD,18}$	0.00(1.00)	*(*)	*(*)	* (*)	* (*)	* (*)
	$Persist_{RD,24}$	* (*)	* (*)	* (*)	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00	*	*	*

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.5.2.2a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate > 4.0mm h^{-1}

Ordinary Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
H Hit Rate	Most Likely	0.74 (0.50)	0.68 (0.50)	0.83 (0.50)	0.74 (0.50)	0.58 (0.47)	0.66 (0.48)
	Prob. Median	0.75 (0.50)	0.74 (0.50)	0.81 (0.50)	0.75 (0.50)	0.57 (0.45)	0.72 (0.52)
	$Persist_{RG,18}$	0.64 (0.50)	0.72 (0.50)	0.72 (0.50)	0.68 (0.50)	0.55 (0.44)	0.55 (0.44)
	$Persist_{RG,24}$	0.62 (0.50)	0.55 (0.50)	0.57 (0.50)	0.62 (0.48)	0.49 (0.48)	0.51 (0.48)
	$Persist_{RD,18}$	0.64 (0.48)	0.55 (0.48)	0.60 (0.49)	0.75 (0.58)	0.72 (0.70)	0.74 (0.67)
	$Persist_{RD,24}$	0.58 (0.48)	0.51 (0.48)	0.60 (0.49)	0.66 (0.56)	0.75 (0.78)	0.74 (0.69)
	Climatology	0.50	0.50	0.50	0.56	0.77	0.69
CSI Critical Success Index	Most Likely	0.56 (0.31)	0.58 (0.30)	0.70 (0.32)	0.63 (0.40)	0.52 (0.42)	0.58 (0.43)
	Prob. Median	0.59 (0.32)	0.55 (0.29)	0.69 (0.34)	0.66 (0.41)	0.50 (0.41)	0.65 (0.47)
	$Persist_{RG,18}$	0.46 (0.32)	0.52 (0.28)	0.52 (0.28)	0.57 (0.41)	0.48 (0.39)	0.45 (0.36)
	$Persist_{RG,24}$	0.41 (0.29)	0.35 (0.31)	0.38 (0.32)	0.50 (0.37)	0.45 (0.44)	0.45 (0.43)
	$Persist_{RD,18}$	0.54 (0.39)	0.47 (0.41)	0.52 (0.43)	0.70 (0.55)	0.71 (0.69)	0.71 (0.66)
	$Persist_{RD,24}$	0.46 (0.37)	0.47 (0.44)	0.53 (0.44)	0.60 (0.51)	0.75 (0.78)	0.72 (0.68)
	Climatology	0.29	0.31	0.33	0.51	0.77	0.68
FAR False Alarm Rate	Most Likely	0.31 (0.55)	0.33 (0.53)	0.16 (0.51)	0.08 (0.32)	0.00 (0.13)	0.00 (0.19)
	Prob. Median	0.30 (0.55)	0.26 (0.53)	0.21 (0.51)	0.07 (0.32)	0.00 (0.13)	0.00 (0.19)
	$Persist_{RG,18}$	0.41 (0.55)	0.27 (0.53)	0.24 (0.51)	0.15 (0.32)	0.00 (0.13)	0.05 (0.19)
	$Persist_{RG,24}$	0.42 (0.55)	0.48 (0.53)	0.44 (0.51)	0.17 (0.32)	0.12 (0.13)	0.16 (0.19)
	$Persist_{RD,18}$	0.44 (0.55)	0.49 (0.53)	0.44 (0.51)	0.21 (0.32)	0.12 (0.13)	0.15 (0.19)
	$Persist_{RD,24}$	0.47 (0.55)	0.51 (0.53)	0.44 (0.51)	0.25 (0.32)	0.15 (0.13)	0.16 (0.19)
	Climatology	0.55	0.53	0.51	0.32	0.13	0.19
POD Probability of Detection	Most Likely	0.75 (0.49)	0.64 (0.45)	0.81 (0.47)	0.67 (0.49)	0.52 (0.45)	0.58 (0.47)
	Prob. Median	0.79 (0.51)	0.68 (0.43)	0.85 (0.53)	0.69 (0.51)	0.50 (0.43)	0.65 (0.53)
	$Persist_{RG,18}$	0.67 (0.51)	0.64 (0.42)	0.62 (0.40)	0.64 (0.51)	0.48 (0.42)	0.47 (0.40)
	$Persist_{RG,24}$	0.58 (0.45)	0.52 (0.47)	0.54 (0.47)	0.56 (0.45)	0.48 (0.47)	0.49 (0.47)
	$Persist_{RD,18}$	0.92 (0.74)	0.84 (0.77)	0.88 (0.77)	0.86 (0.74)	0.78 (0.77)	0.81 (0.77)
	$Persist_{RD,24}$	0.79 (0.68)	0.92 (0.89)	0.92 (0.81)	0.75 (1.68)	0.87 (0.89)	0.84 (0.81)
	Climatology	0.45	0.47	0.49	0.68	0.87	0.81
B Bias Ratio	Most Likely	1.08 (1.08)	0.96 (0.96)	0.96 (0.96)	0.72 (0.72)	0.52 (0.52)	0.58 (0.58)
	Prob. Median	1.12 (1.12)	0.92 (0.92)	1.08 (1.08)	0.75 (0.75)	0.50 (0.50)	0.65 (0.65)
	$Persist_{RG,18}$	1.12 (1.12)	0.88 (0.88)	0.81 (0.81)	0.75 (0.75)	0.48 (0.48)	0.49 (0.49)
	$Persist_{RG,24}$	1.00 (1.00)	1.00 (1.00)	0.96 (0.96)	0.67 (0.67)	0.54 (0.54)	0.58 (0.85)
	$Persist_{RD,18}$	1.62 (1.62)	1.64 (1.64)	1.58 (1.58)	1.08 (1.08)	0.89 (0.89)	0.95 (0.95)
	$Persist_{RD,24}$	1.50 (1.50)	1.88 (1.88)	1.65 (1.65)	1.00 (1.00)	1.02 (1.02)	1.00 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.2b Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate > 4.0mm h^{-1} Relative Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
HSS Heidke Skill Score	Most Likely	0.47 (0.00)	0.36 (0.00)	0.66 (0.00)	0.48 (0.00)	0.22 (0.00)	0.34 (0.00)
	Prob. Median	0.51 (0.00)	0.47 (0.00)	0.62 (0.00)	0.51 (0.00)	0.11 (0.00)	0.41 (0.00)
	$Persist_{RG,18}$	0.28 (0.00)	0.43 (0.00)	0.43 (0.00)	0.35 (0.00)	0.19 (0.00)	0.20 (0.00)
	$Persist_{RG,24}$	0.34 (0.00)	0.09 (0.00)	0.13 (0.00)	0.27 (0.00)	0.02 (0.00)	0.05 (0.00)
	$Persist_{RD,18}$	0.31 (0.00)	0.12 (0.00)	0.22 (0.00)	0.41 (0.00)	0.05 (0.00)	0.20 (0.00)
	$Persist_{RD,24}$	0.20 (0.00)	0.06 (0.00)	0.22 (0.00)	0.22 (0.00)	-0.14(0.00)	0.14 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.47 (0.00)	0.35 (0.00)	0.66 (0.00)	0.55 (0.00)	0.52 (0.00)	0.58 (0.00)
	Prob. Median	0.52 (0.00)	0.47 (0.00)	0.62 (0.00)	0.58 (0.00)	0.50 (0.00)	0.65 (0.00)
	$Persist_{RG,18}$	0.29 (0.00)	0.43 (0.00)	0.43 (0.00)	0.40 (0.00)	0.48 (0.00)	0.37 (0.00)
	$Persist_{RG,24}$	0.24 (0.00)	0.09 (0.00)	0.13 (0.00)	0.32 (0.00)	0.05 (0.00)	0.09 (0.00)
	$Persist_{RD,18}$	0.33 (0.00)	0.13 (0.00)	0.22 (0.00)	0.39 (0.00)	-0.13(0.00)	0.21 (0.00)
	$Persist_{RD,24}$	0.21 (0.00)	0.06 (0.00)	0.22 (0.00)	0.22 (0.00)	0.22 (0.00)	0.14 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.31 (0.00)	0.22 (0.00)	0.49 (0.00)	0.31 (0.00)	0.13 (0.00)	0.21 (0.00)
	Prob. Median	0.34 (0.00)	0.31 (0.00)	0.45 (0.00)	0.34 (0.00)	0.12 (0.00)	0.26 (0.00)
	$Persist_{RG,18}$	0.17 (0.00)	0.27 (0.00)	0.28 (0.00)	0.22 (0.00)	0.11 (0.00)	0.11 (0.00)
	$Persist_{RG,24}$	0.14 (0.00)	0.05 (0.00)	0.07 (0.00)	0.16 (0.00)	0.01 (0.00)	0.03 (0.00)
	$Persist_{RD,18}$	0.19 (0.00)	0.06 (0.00)	0.12 (0.00)	0.26 (0.00)	0.03 (0.00)	0.11 (0.00)
	$Persist_{RD,24}$	0.11 (0.00)	0.03 (0.00)	0.12 (0.00)	0.12 (0.00)	-0.06(0.00)	0.07 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	2.72 (1.00)	2.24 (1.00)	5.45 (1.00)	5.67 (1.00)	* (1.00)	* (1.00)
	Prob. Median	2.87 (1.00)	3.17 (1.00)	3.81 (1.00)	5.90 (1.00)	* (1.00)	* (1.00)
	$Persist_{RG,18}$	1.76 (1.00)	2.99 (1.00)	3.32 (1.00)	2.72 (1.00)	* (1.00)	4.65 (1.00)
	$Persist_{RG,24}$	1.69 (1.00)	1.21 (1.00)	1.32 (1.00)	2.36 (1.00)	1.12 (1.00)	1.22 (1.00)
	$Persist_{RD,18}$	1.56 (1.00)	1.18 (1.00)	1.33 (1.00)	1.83 (1.00)	1.10 (1.00)	1.36 (1.00)
	$Persist_{RD,24}$	1.35 (1.00)	1.07 (1.00)	1.31 (1.00)	1.42 (1.00)	0.87 (1.00)	1.20 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	2.90 (1.00)	1.98 (1.00)	4.43 (1.00)	2.65 (1.00)	2.09 (1.00)	2.39 (1.00)
	Prob. Median	3.48 (1.00)	2.46 (1.00)	5.06 (1.00)	2.89 (1.00)	2.00 (1.00)	2.87 (1.00)
	$Persist_{RG,18}$	1.86 (1.00)	2.18 (1.00)	2.12 (1.00)	2.12 (1.00)	1.92 (1.00)	1.68 (1.00)
	$Persist_{RG,24}$	1.57 (1.00)	1.19 (1.00)	1.28 (1.00)	1.72 (1.00)	1.10 (1.00)	1.17 (1.00)
	$Persist_{RD,18}$	4.97 (1.00)	1.79 (1.00)	2.89 (1.00)	3.81 (1.00)	1.31 (1.00)	2.15 (1.00)
	$Persist_{RD,24}$	1.99 (1.00)	1.79 (1.00)	3.85 (1.00)	1.88 (1.00)	0.00 (1.00)	1.84 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
θ Odds Ratio	Most Likely	7.87 (1.00)	4.44 (1.00)	24.1 (1.00)	15.0 (1.00)	* (1.00)	* (1.00)
	Prob. Median	9.98 (1.00)	7.79 (1.00)	19.2 (1.00)	17.0 (1.00)	* (1.00)	* (1.00)
	$Persist_{RG,18}$	3.27 (1.00)	6.52 (1.00)	7.04 (1.00)	5.75 (1.00)	* (1.00)	7.83 (1.00)
	$Persist_{RG,24}$	2.66 (1.00)	1.44 (1.00)	1.70 (1.00)	4.06 (1.00)	1.22 (1.00)	1.43 (1.00)
	$Persist_{RD,18}$	7.76 (1.00)	2.10 (1.00)	3.83 (1.00)	6.98 (1.00)	1.44 (1.00)	2.92 (1.00)
	$Persist_{RD,24}$	2.68 (1.00)	1.92 (1.00)	5.05 (1.00)	2.67 (1.00)	0.00 (1.00)	2.20 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.5.2.3a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate >12.0mm h^{-1}

Ordinary Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
H Hit Rate	Most Likely	0.81 (0.69)	0.74 (0.64)	0.83 (0.68)	0.64 (0.48)	0.53 (0.41)	0.57 (0.42)
	Prob. Median	0.77 (0.72)	0.72 (0.65)	0.81 (0.69)	0.60 (0.48)	0.51 (0.40)	0.55 (0.42)
	$Persist_{RG,18}$	0.75 (0.59)	0.68 (0.60)	0.77 (0.69)	0.62 (0.49)	0.55 (0.43)	0.51 (0.42)
	$Persist_{RG,24}$	0.77 (0.67)	0.58 (0.59)	0.74 (0.67)	0.53 (0.48)	0.53 (0.44)	0.51 (0.43)
	$Persist_{RD,18}$	0.68 (0.51)	0.64 (0.49)	0.66 (0.49)	0.66 (0.50)	0.66 (0.51)	0.66 (0.50)
	$Persist_{RD,24}$	0.53 (0.48)	0.51 (0.43)	0.47 (0.43)	0.62 (0.40)	0.64 (0.55)	0.62 (0.53)
	Climatology	0.67	0.59	0.67	0.50	0.54	0.53
CSI Critical Success Index	Most Likely	0.33 (0.10)	0.26 (0.12)	0.40 (0.11)	0.32 (0.15)	0.26 (0.16)	0.30 (0.17)
	Prob. Median	0.20 (0.09)	0.21 (0.11)	0.33 (0.10)	0.25 (0.12)	0.24 (0.14)	0.27 (0.15)
	$Persist_{RG,18}$	0.38 (0.15)	0.26 (0.16)	0.25 (0.10)	0.39 (0.26)	0.33 (0.23)	0.24 (0.15)
	$Persist_{RG,24}$	0.29 (0.12)	0.15 (0.16)	0.22 (0.12)	0.22 (0.18)	0.32 (0.24)	0.26 (0.18)
	$Persist_{RD,18}$	0.37 (0.17)	0.39 (0.23)	0.36 (0.17)	0.50 (0.34)	0.55 (0.41)	0.54 (0.39)
	$Persist_{RD,24}$	0.22 (0.18)	0.32 (0.25)	0.22 (0.18)	0.47 (0.36)	0.57 (0.48)	0.53 (0.45)
	Climatology	0.12	0.16	0.12	0.36	0.47	0.45
FAR False Alarm Rate	Most Likely	0.44 (0.79)	0.44 (0.72)	0.40 (0.79)	0.00 (0.47)	0.00 (0.36)	0.00 (0.38)
	Prob. Median	0.57 (0.79)	0.50 (0.72)	0.44 (0.79)	0.00 (0.47)	0.00 (0.36)	0.00 (0.38)
	$Persist_{RG,18}$	0.56 (0.79)	0.57 (0.72)	0.56 (0.79)	0.28 (0.47)	0.14 (0.36)	0.11 (0.38)
	$Persist_{RG,24}$	0.55 (0.79)	0.73 (0.72)	0.64 (0.79)	0.36 (0.47)	0.20 (0.36)	0.18 (0.38)
	$Persist_{RD,18}$	0.62 (0.79)	0.57 (0.72)	0.63 (0.79)	0.31 (0.47)	0.21 (0.36)	0.22 (0.38)
	$Persist_{RD,24}$	0.75 (0.79)	0.66 (0.72)	0.76 (0.79)	0.36 (0.47)	0.29 (0.36)	0.30 (0.38)
	Climatology	0.79	0.72	0.79	0.47	0.36	0.38
POD Probability of Detection	Most Likely	0.45 (0.17)	0.33 (0.17)	0.55 (0.19)	0.32 (0.17)	0.26 (0.17)	0.30 (0.19)
	Prob. Median	0.27 (0.13)	0.27 (0.15)	0.45 (0.17)	0.25 (0.13)	0.24 (0.15)	0.27 (0.17)
	$Persist_{RG,18}$	0.73 (0.34)	0.40 (0.26)	0.36 (0.17)	0.46 (0.34)	0.35 (0.26)	0.24 (0.17)
	$Persist_{RG,24}$	0.45 (0.21)	0.27 (0.28)	0.36 (0.21)	0.25 (0.21)	0.35 (0.28)	0.27 (0.21)
	$Persist_{RD,18}$	0.91 (0.49)	0.80 (0.53)	0.91 (0.51)	0.64 (0.49)	0.65 (0.53)	0.64 (0.51)
	$Persist_{RD,24}$	0.64 (0.53)	0.80 (0.66)	0.73 (0.62)	0.64 (0.53)	0.74 (0.66)	0.70 (0.62)
	Climatology	0.45	0.47	0.49	0.68	0.64	0.62
B Bias Ratio	Most Likely	0.82 (0.82)	0.60 (0.60)	0.91 (0.91)	0.32 (0.32)	0.26 (0.26)	0.30 (0.30)
	Prob. Median	0.64 (0.64)	0.53 (0.53)	0.82 (0.82)	0.25 (0.25)	0.24 (0.24)	0.27 (0.27)
	$Persist_{RG,18}$	1.64 (1.64)	0.93 (0.93)	0.82 (0.82)	0.64 (0.64)	0.41 (0.41)	0.27 (0.27)
	$Persist_{RG,24}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.39 (0.39)	0.44 (0.44)	0.33 (0.33)
	$Persist_{RD,18}$	2.36 (2.36)	1.87 (1.87)	2.45 (2.45)	0.93 (0.93)	0.82 (0.82)	0.82 (0.82)
	$Persist_{RD,24}$	2.55 (2.55)	2.33 (2.33)	3.00 (3.00)	1.00 (1.00)	1.03 (1.03)	1.00 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.3b Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate > 12.0mm h^{-1} Relative Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
HSS Heidke Skill Score	Most Likely	0.39 (0.00)	0.26 (0.00)	0.47 (0.00)	0.31 (0.00)	0.21 (0.00)	0.25 (0.00)
	Prob. Median	0.20 (0.00)	0.19 (0.00)	0.39 (0.00)	0.24 (0.00)	0.18 (0.00)	0.22 (0.00)
	$Persist_{RG,18}$	0.40 (0.00)	0.19 (0.00)	0.26 (0.00)	0.26 (0.00)	0.20 (0.00)	0.16 (0.00)
	$Persist_{RG,24}$	0.31 (0.00)	-0.02(0.00)	0.20 (0.00)	0.09 (0.00)	0.16 (0.00)	0.14 (0.00)
	$Persist_{RD,18}$	0.35 (0.00)	0.30 (0.00)	0.33 (0.00)	0.32 (0.00)	0.31 (0.00)	0.32 (0.00)
	$Persist_{RD,24}$	0.09 (0.00)	0.14 (0.00)	0.08 (0.00)	0.24 (0.00)	0.21 (0.00)	0.20 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.36 (0.00)	0.35 (0.00)	0.45 (0.00)	0.32 (0.00)	0.26 (0.00)	0.30 (0.00)
	Prob. Median	0.18 (0.00)	0.47 (0.00)	0.36 (0.00)	0.25 (0.00)	0.24 (0.00)	0.27 (0.00)
	$Persist_{RG,18}$	0.49 (0.00)	0.43 (0.00)	0.24 (0.00)	0.26 (0.00)	0.25 (0.00)	0.19 (0.00)
	$Persist_{RG,24}$	0.31 (0.00)	0.09 (0.00)	0.20 (0.00)	0.09 (0.00)	0.20 (0.00)	0.17 (0.00)
	$Persist_{RD,18}$	0.53 (0.00)	0.13 (0.00)	0.50 (0.00)	0.32 (0.00)	0.33 (0.00)	0.34 (0.00)
	$Persist_{RD,24}$	0.14 (0.00)	0.06 (0.00)	0.13 (0.00)	0.24 (0.00)	0.21 (0.00)	0.20 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.24 (0.00)	0.15 (0.00)	0.30 (0.00)	0.18 (0.00)	0.11 (0.00)	0.14 (0.00)
	Prob. Median	0.11 (0.00)	0.10 (0.00)	0.24 (0.00)	0.14 (0.00)	0.10 (0.00)	0.12 (0.00)
	$Persist_{RG,18}$	0.25 (0.00)	0.11 (0.00)	0.15 (0.00)	0.15 (0.00)	0.11 (0.00)	0.08 (0.00)
	$Persist_{RG,24}$	0.18 (0.00)	-0.01(0.00)	0.11 (0.00)	0.05 (0.00)	0.09 (0.00)	0.08 (0.00)
	$Persist_{RD,18}$	0.21 (0.00)	0.18 (0.00)	0.20 (0.00)	0.19 (0.00)	0.18 (0.00)	0.19 (0.00)
	$Persist_{RD,24}$	0.05 (0.00)	0.07 (0.00)	0.04 (0.00)	0.14 (0.00)	0.12 (0.00)	0.11 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	4.77 (1.00)	3.17 (1.00)	5.73 (1.00)	* (1.00)	* (1.00)	* (1.00)
	Prob. Median	2.86 (1.00)	2.53 (1.00)	4.77 (1.00)	* (1.00)	* (1.00)	* (1.00)
	$Persist_{RG,18}$	3.05 (1.00)	1.90 (1.00)	3.05 (1.00)	2.32 (1.00)	3.35 (1.00)	4.85 (1.00)
	$Persist_{RG,24}$	3.18 (1.00)	0.92 (1.00)	2.18 (1.00)	1.56 (1.00)	2.24 (1.00)	2.73 (1.00)
	$Persist_{RD,18}$	2.39 (1.00)	1.90 (1.00)	2.25 (1.00)	2.01 (1.00)	2.05 (1.00)	2.12 (1.00)
	$Persist_{RD,24}$	1.27 (1.00)	1.32 (1.00)	1.22 (1.00)	1.61 (1.00)	1.40 (1.00)	1.39 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	1.66 (1.00)	1.34 (1.00)	1.99 (1.00)	1.47 (1.00)	1.36 (1.00)	1.43 (1.00)
	Prob. Median	1.24 (1.00)	1.22 (1.00)	1.66 (1.00)	1.33 (1.00)	1.31 (1.00)	1.38 (1.00)
	$Persist_{RG,18}$	2.79 (1.00)	1.32 (1.00)	1.38 (1.00)	1.49 (1.00)	1.38 (1.00)	1.25 (1.00)
	$Persist_{RG,24}$	1.57 (1.00)	0.97 (1.00)	1.31 (1.00)	1.12 (1.00)	1.30 (1.00)	1.24 (1.00)
	$Persist_{RD,18}$	6.81 (1.00)	2.89 (1.00)	6.55 (1.00)	1.90 (1.00)	1.94 (1.00)	1.92 (1.00)
	$Persist_{RD,24}$	1.38 (1.00)	1.97 (1.00)	1.48 (1.00)	1.68 (1.00)	1.79 (1.00)	1.65 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
θ Odds Ratio	Most Likely	7.92 (1.00)	4.25 (1.00)	11.4 (1.00)	* (1.00)	* (1.00)	* (1.00)
	Prob. Median	3.56 (1.00)	3.09 (1.00)	7.92 (1.00)	* (1.00)	* (1.00)	* (1.00)
	$Persist_{RG,18}$	8.53 (1.00)	2.50 (1.00)	4.32 (1.00)	3.47 (1.00)	4.64 (1.00)	6.08 (1.00)
	$Persist_{RG,24}$	5.0 (1.00)	0.89 (1.00)	2.86 (1.00)	1.75 (1.00)	2.91 (1.00)	3.38 (1.00)
	$Persist_{RD,18}$	16.2 (1.00)	5.50 (1.00)	14.7 (1.00)	3.83 (1.00)	3.97 (1.00)	4.08 (1.00)
	$Persist_{RD,24}$	1.75 (1.00)	2.61 (1.00)	5.05 (1.00)	2.70 (1.00)	2.50 (1.00)	2.30 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.5.2.4a Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate > 25.0mm h^{-1}

Ordinary Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
H Hit Rate	Most Likely	0.85 (0.86)	0.87 (0.88)	0.89 (0.86)	0.64 (0.63)	0.64 (0.63)	0.66 (0.61)
	Prob. Median	0.83 (0.84)	0.85 (0.86)	0.85 (0.86)	0.66 (0.63)	0.66 (0.63)	0.66 (0.61)
	$Persist_{RG,18}$	0.79 (0.73)	0.87 (0.78)	0.89 (0.86)	0.66 (0.58)	0.68 (0.60)	0.66 (0.61)
	$Persist_{RG,24}$	0.81 (0.80)	0.81 (0.83)	0.89 (0.80)	0.64 (0.61)	0.58 (0.61)	0.74 (0.59)
	$Persist_{RD,18}$	0.62 (0.54)	0.58 (0.55)	0.77 (0.62)	0.68 (0.51)	0.62 (0.52)	0.74 (0.54)
	$Persist_{RD,24}$	0.60 (0.61)	0.64 (0.60)	0.66 (0.59)	0.58 (0.54)	0.68 (0.53)	0.74 (0.53)
	Climatology	0.80	0.83	0.80	0.54	0.54	0.53
CSI Critical Success Index	Most Likely	0.00 (0.03)	0.00 (0.03)	0.14 (0.03)	0.05 (0.04)	0.05 (0.04)	0.10 (0.04)
	Prob. Median	0.00 (0.04)	0.00 (0.04)	0.00 (0.03)	0.10 (0.05)	0.10 (0.05)	0.10 (0.04)
	$Persist_{RG,18}$	0.21 (0.08)	0.30 (0.06)	0.14 (0.03)	0.25 (0.15)	0.23 (0.12)	0.10 (0.04)
	$Persist_{RG,24}$	0.09 (0.06)	0.00 (0.05)	0.33 (0.06)	0.14 (0.09)	0.04 (0.08)	0.30 (0.10)
	$Persist_{RD,18}$	0.20 (0.10)	0.12 (0.08)	0.33 (0.09)	0.43 (0.25)	0.35 (0.24)	0.46 (0.22)
	$Persist_{RD,24}$	0.09 (0.09)	0.14 (0.08)	0.18 (0.10)	0.27 (0.22)	0.39 (0.23)	0.48 (0.23)
	Climatology	0.06	0.05	0.16	0.22	0.22	0.23
FAR False Alarm Rate	Most Likely	1.00 (0.89)	1.00 (0.91)	0.50 (0.89)	0.50 (0.64)	0.50 (0.64)	0.00 (0.62)
	Prob. Median	1.00 (0.89)	1.00 (0.91)	1.00 (0.89)	0.33 (0.64)	0.33 (0.64)	0.00 (0.62)
	$Persist_{RG,18}$	0.73 (0.89)	0.62 (0.91)	0.50 (0.89)	0.45 (0.64)	0.38 (0.64)	0.00 (0.62)
	$Persist_{RG,24}$	0.83 (0.89)	1.00 (0.91)	0.50 (0.89)	0.50 (0.64)	0.80 (0.64)	0.00 (0.62)
	$Persist_{RD,18}$	0.79 (0.89)	0.87 (0.91)	0.67 (0.89)	0.46 (0.64)	0.52 (0.64)	0.33 (0.62)
	$Persist_{RD,24}$	0.89 (0.89)	0.85 (0.91)	0.80 (0.89)	0.58 (0.64)	0.45 (0.64)	0.35 (0.62)
	Climatology	0.89	0.91	0.89	0.64	0.64	0.62
POD Probability of Detection	Most Likely	0.00 (0.04)	0.00 (0.04)	0.17 (0.04)	0.05 (0.04)	0.05 (0.04)	0.10 (0.19)
	Prob. Median	0.00 (0.06)	0.00 (0.06)	0.00 (0.04)	0.11 (0.06)	0.11 (0.06)	0.10 (0.17)
	$Persist_{RG,18}$	0.50 (0.21)	0.60 (0.15)	0.17 (0.04)	0.32 (0.21)	0.26 (0.15)	0.10 (0.17)
	$Persist_{RG,24}$	0.17 (0.11)	0.00 (0.09)	0.50 (0.11)	0.16 (0.11)	0.05 (0.09)	0.30 (0.21)
	$Persist_{RD,18}$	0.83 (0.45)	0.60 (0.43)	1.00 (0.34)	0.68 (0.45)	0.58 (0.43)	0.60 (0.51)
	$Persist_{RD,24}$	0.33 (0.36)	0.60 (0.38)	0.67 (0.38)	0.42 (0.36)	0.58 (0.38)	0.65 (0.62)
	Climatology	0.11	0.09	0.11	0.36	0.36	0.38
B Bias Ratio	Most Likely	0.33 (0.33)	0.40 (0.40)	0.33 (0.33)	0.11 (0.11)	0.11 (0.21)	0.10 (0.10)
	Prob. Median	0.50 (0.50)	0.60 (0.60)	0.33 (0.33)	0.16 (0.16)	0.16 (0.16)	0.10 (0.10)
	$Persist_{RG,18}$	1.83 (1.83)	1.60 (1.60)	0.33 (0.33)	0.58 (0.58)	0.42 (0.42)	0.10 (0.10)
	$Persist_{RG,24}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	0.32 (0.32)	0.26 (0.26)	0.30 (0.30)
	$Persist_{RD,18}$	4.00 (4.00)	4.60 (4.60)	3.00 (3.00)	1.26 (0.26)	1.21 (1.21)	0.90 (0.90)
	$Persist_{RD,24}$	3.17 (3.17)	4.00 (4.00)	3.33 (3.33)	1.00 (1.00)	1.05 (1.05)	1.00 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.4.3.2.4b Categorical assessment measures for Evening Update forecasts in the Thames Region. Rainfall Rate > 25.0mm h^{-1} Relative Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
HSS Heidke Skill Score	Most Likely	-0.06(0.00)	-0.06(0.00)	0.20 (0.00)	0.03 (0.00)	0.03 (0.00)	0.12 (0.00)
	Prob. Median	-0.08(0.00)	-0.08(0.00)	-0.06(0.00)	0.09 (0.00)	0.09 (0.00)	0.12 (0.00)
	$Persist_{RG,18}$	0.24 (0.00)	0.39 (0.00)	0.20 (0.00)	0.19 (0.00)	0.20 (0.00)	0.12 (0.00)
	$Persist_{RG,24}$	0.06 (0.00)	-0.10(0.00)	0.44 (0.00)	0.08 (0.00)	-0.08(0.00)	0.35 (0.00)
	$Persist_{RD,18}$	0.19 (0.00)	0.07 (0.00)	0.40 (0.00)	0.34 (0.00)	0.22 (0.00)	0.43 (0.00)
	$Persist_{RD,24}$	-0.01(0.00)	0.10 (0.00)	0.16 (0.00)	0.10 (0.00)	0.31 (0.00)	0.44 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	-0.04(0.00)	-0.04(0.00)	0.15 (0.00)	0.02 (0.00)	0.02 (0.00)	0.10 (0.00)
	Prob. Median	-0.06(0.00)	-0.06(0.00)	-0.04(0.00)	0.08 (0.00)	0.08 (0.00)	0.10 (0.00)
	$Persist_{RG,18}$	0.33 (0.00)	0.50 (0.00)	0.15 (0.00)	0.17 (0.00)	0.17 (0.00)	0.10 (0.00)
	$Persist_{RG,24}$	0.06 (0.00)	-0.10(0.00)	0.44 (0.00)	0.07 (0.00)	-0.07(0.00)	0.30 (0.00)
	$Persist_{RD,18}$	0.43 (0.00)	0.18 (0.00)	0.74 (0.00)	0.36 (0.00)	0.23 (0.00)	0.42 (0.00)
	$Persist_{RD,24}$	-0.03(0.00)	0.25 (0.00)	0.33 (0.00)	0.10 (0.00)	0.31 (0.00)	0.44 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	-0.03(0.00)	-0.03(0.00)	0.11 (0.00)	0.01 (0.00)	0.01 (0.00)	0.06 (0.00)
	Prob. Median	-0.04(0.00)	-0.04(0.00)	-0.03(0.00)	0.05 (0.00)	0.05 (0.00)	0.06 (0.00)
	$Persist_{RG,18}$	0.14 (0.00)	0.24 (0.00)	0.11 (0.00)	0.10 (0.00)	0.11 (0.00)	0.06 (0.00)
	$Persist_{RG,24}$	0.03 (0.00)	-0.05(0.00)	0.28 (0.00)	0.04 (0.00)	-0.04(0.00)	0.21 (0.00)
	$Persist_{RD,18}$	0.10 (0.00)	0.04 (0.00)	0.25 (0.00)	0.21 (0.00)	0.12 (0.00)	0.27 (0.00)
	$Persist_{RD,24}$	-0.01(0.00)	0.06 (0.00)	0.09 (0.00)	0.05 (0.00)	0.18 (0.00)	0.28 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	0.00 (1.00)	0.00 (1.00)	7.83 (1.00)	1.79 (1.00)	1.79 (1.00)	* (1.00)
	Prob. Median	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	3.58 (1.00)	3.58 (1.00)	* (1.00)
	$Persist_{RG,18}$	2.94 (1.00)	5.76 (1.00)	7.83 (1.00)	2.15 (1.00)	2.98 (1.00)	* (1.00)
	$Persist_{RG,24}$	1.57 (1.00)	0.00 (1.00)	7.83 (1.00)	1.79 (1.00)	0.45 (1.00)	* (1.00)
	$Persist_{RD,18}$	2.06 (1.00)	1.44 (1.00)	3.92 (1.00)	2.11 (1.00)	1.64 (1.00)	3.30 (1.00)
	$Persist_{RD,24}$	0.92 (1.00)	1.69 (1.00)	1.96 (1.00)	1.30 (1.00)	2.19 (1.00)	3.06 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	0.96 (1.00)	0.96 (1.00)	1.17 (1.00)	1.02 (1.00)	1.02 (1.00)	1.11 (1.00)
	Prob. Median	0.94 (1.00)	0.94 (1.00)	0.96 (1.00)	1.08 (1.00)	1.08 (1.00)	1.11 (1.00)
	$Persist_{RG,18}$	1.66 (1.00)	2.24 (1.00)	1.17 (1.00)	1.25 (1.00)	1.24 (1.00)	1.11 (1.00)
	$Persist_{RG,24}$	1.07 (1.00)	0.90 (1.00)	1.87 (1.00)	1.08 (1.00)	0.93 (1.00)	1.43 (1.00)
	$Persist_{RD,18}$	3.57 (1.00)	1.46 (1.00)	* (1.00)	2.14 (1.00)	1.54 (1.00)	2.05 (1.00)
	$Persist_{RD,24}$	0.96 (1.00)	1.61 (1.00)	1.98 (1.00)	1.17 (1.00)	1.75 (1.00)	2.25 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
θ Odds Ratio	Most Likely	0.00 (1.00)	0.00 (1.00)	9.20 (1.00)	1.83 (1.00)	1.83 (1.00)	* (1.00)
	Prob. Median	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	3.88 (1.00)	3.88 (1.00)	* (1.00)
	$Persist_{RG,18}$	4.88 (1.00)	12.9 (1.00)	9.20 (1.00)	2.68 (1.00)	3.69 (1.00)	* (1.00)
	$Persist_{RG,24}$	1.68 (1.00)	0.00 (1.00)	14.7 (1.00)	1.94 (1.00)	0.42 (1.00)	* (1.00)
	$Persist_{RD,18}$	7.37 (1.00)	2.10 (1.00)	* (1.00)	4.53 (1.00)	2.52 (1.00)	6.75 (1.00)
	$Persist_{RD,24}$	0.88 (1.00)	2.74 (1.00)	3.88 (1.00)	1.52 (1.00)	3.82 (1.00)	6.90 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

5.4.6 Assessment of Probability Forecasts of Rates

This section outlines an analysis which assesses the performance of the probability forecasts for maximum rainfall rates which are part of the Evening Update forecasts. The analysis is directly comparable to that employed for probability forecasts of maximum rainfall amounts reported in Section 5.5.4.

As for the case of rainfall amounts, two sets of forecasts are used for comparison against the probability forecasts contained in the Evening Updates. As outlined in Section 5.4.1 and Table 5.4.1.2, one set of probability forecasts have been created from the set of single-valued forecasts by treating these as expressing complete certainty in the quoted value, and a second set has been created by taking each single-valued forecast and attaching a somewhat arbitrary uncertainty-band: when the forecast maximum rate is moderately large, this band extends from 0 up to twice the central forecast rate. The specification of this uncertainty band has not been subjected to detailed consideration and is simply put forward for comparison against the performance of the Evening Update probability forecasts.

As for the other analyses of rainfall rates, two different versions of ground-truth are available, and results are given here for both.

The results of the analysis of the probability forecasts are given in Table 5.4.6.1. The upper part of the table relates to the performance of the single-valued forecasts when treated as expressing absolute certainty. Values here for the "certain" forecasts are identical to those for the Mean Absolute Error given in Table 5.4.5.1.1 and they are repeated here for comparison with the results of the other forecasts which do contain uncertainty. The lower part of the Table gives the Continuous Brier Score for the constructed probability forecasts and for the Evening Updates' probability forecasts. It can be seen that including the uncertainty band with the single-valued forecasts has always decreased the performance measure in these cases. However, note that adding uncertainty of greater amounts would eventually lead to an increase in the score.

As for the analysis of the single-valued forecasts in Section 5.4.5.1, Table 5.4.6.1 again shows that the forecasts (and in particular the probability forecast) contained in the Evening Updates are a better match to the outcomes derived from the raingauge network than they are to those obtained from the Nimrod radar-rainfall source.

In Table 5.4.6.1, the results for the Evening Updates' probability forecasts are again somewhat disappointing in comparison with those for the constructed probability forecasts, particularly when considering the probability forecast obtained from the "Most Likely" forecast by adding a simple uncertainty band. It seems that the probability forecasts contained in the Evening Updates are not much better, if at all better, than could be obtained by a simple uncertainty band centred about the main forecast-value.

Tables 5.4.6.2 and 5.4.6.3 relate directly to the question of whether there is enough evidence in the test dataset to distinguish between the performances of the different types of probability forecast. Taking the 'most likely' forecast ("Most Likely") from the Evening Updates, with the addition of either zero or 100% uncertainty, as a "base forecast", Table 5.4.6.2 considers each of the other forecast sources in turn and asks

how much evidence there is that the "base forecast" is better than the candidate forecast. In Table 5.4.6.3 the "base forecast" is the probability forecast contained in the Evening Updates. The values in these tables are the standardised differences discussed earlier in Section 4.3.3, and positive values indicate that the "base forecast" has a better performance, as measured by the Continuous Brier Score, than the candidate. If the candidate forecast had a better performance, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the Continuous Brier Scores for the two forecast sources will turn out to be in the order indicated. For the purposes here, a standardised difference outside the range ± 2 units indicates fairly strong evidence that one forecast source really is better than another.

The results shown in Tables 5.4.6.2 and 5.4.6.3 can be interpreted as follows for the case of the raingauge-based ground-truth Firstly, both the operational probability forecast and the probability forecast derived by adding a 100% uncertainty band to the 'Most Likely' forecast are better than the probability forecasts constructed by attaching 100% uncertainty bands to the persistence forecasts derived from radar or the constant-valued forecasts. The raingauge-based persistence forecasts based on the immediately preceding 18-hour period are close to have been shown to be worse than the forecasts from the Evening Updates. However, the size of the test dataset is too small to allow a clear distinction to be made between the operational probability forecast from the Evening Updates and the simple type of probability forecast derived by adding a 100% uncertainty band to the 'Most Likely' forecast that is contained in the Evening Updates: but the results here favour the latter in two of the three sub-areas. In the case of the radar-based ground-truth, the extent of the mismatch in the values produced as forecasts in the Evening Update and those actually observed in the radar data is such that none of the probability forecasts are good and they are not clearly distinguishable. The probability forecast derived from the constant-valued forecast of 0 mm h^{-1} is clearly worse than the forecasts contained in, or derived from the Evening Update, but a constant-valued forecast consisting of a uniform distribution over the range 0-20 mm h^{-1} (i.e. "$Const_{10mm/hr}$" with 100% error band) appears to be competitive with the operational forecasts (but neither is good at forecasting the radar-derived maximum rainfall rates).

The conclusions here about the probability forecasts in the Evening Updates needs to be tempered by the same consideration as outlined at the end of Section 5.4.4. Specifically, that better results might have been obtained for the probability forecasts in the Evening Updates if they had been produced at a finer resolution.

Table 5.4.6.1 Assessment measures for Probability Forecasts associated with Evening Update forecasts in the Thames Region. (Rainfall Rates)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error (mm/hr)	*(certain)*						
	Most Likely	6.84	7.68	5.28	26.3	24.7	25.7
	Prob. Median	7.03	7.41	5.61	26.9	25.0	26.2
	$Persist_{RG,18}$	9.45	8.77	6.55	24.4	23.4	26.3
	$Persist_{RG,24}$	9.31	11.2	7.47	27.6	27.4	26.3
	$Persist_{RD,18}$	28.5	29.6	25.2	28.5	26.3	22.7
	$Persist_{RD,24}$	28.9	26.4	27.3	37.1	26.2	28.8
	$Const_{0mm/hr}$	8.85	8.78	8.38	32.1	30.6	32.6
	$Const_{10mm/hr}$	9.28	9.02	8.17	27.8	23.8	26.7
Continuous Brier Score (mm/hr)	*(100% error)*						
	Most Likely	5.11	6.03	3.84	24.7	22.6	23.7
	Prob. Median	5.40	5.87	4.17	25.5	23.1	23.9
	$Persist_{RG,18}$	6.65	6.52	5.22	21.8	21.1	24.2
	$Persist_{RG,24}$	6.87	8.63	6.14	25.5	25.2	24.1
	$Persist_{RD,18}$	17.3	18.3	15.2	21.2	19.1	18.6
	$Persist_{RD,24}$	18.2	16.6	16.9	28.0	18.9	20.8
	$Const_{0mm/hr}$	8.50	8.47	8.01	31.7	30.1	32.1
	$Const_{10mm/hr}$	6.56	6.29	5.84	24.8	21.2	23.9
	(operational)						
	Prob. Forecast	5.06	6.61	4.03	23.9	21.7	22.9

Table 5.4.6.2 Comparison of forecast sources: Standardised Differences for Assessment Measures for Probability Forecasts associated with Evening Update forecasts in the Thames Region. (Rainfall Rates) (In this Table, the base forecast is "Most Likely" with either zero or 100% error)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error (mm/hr)	*(certain)*						
	Prob. Median	0.56	-0.68	0.73	1.65	0.81	0.98
	$Persist_{RG,18}$	1.56	1.19	1.27	-0.96	-0.69	0.51
	$Persist_{RG,24}$	1.89	2.45	1.88	1.02	2.02	0.43
	$Persist_{RD,18}$	4.26	4.84	3.89	0.38	0.56	-0.61
	$Persist_{RD,24}$	4.09	4.58	4.33	2.36	0.39	0.55
	$Const_{0mm/hr}$	1.94	1.32	2.97	5.87	6.07	6.26
	$Const_{10mm/hr}$	2.50	1.31	2.98	1.64	-0.82	0.91
Continuous Brier Score (mm/hr)	*(100% error)*						
	Prob. Median	1.05	-0.59	0.94	1.83	1.33	0.58
	$Persist_{RG,18}$	1.42	0.82	2.06	-1.39	-0.74	0.41
	$Persist_{RG,24}$	1.72	2.75	2.59	0.53	2.02	0.25
	$Persist_{RD,18}$	3.98	4.68	3.79	-0.81	-1.32	-1.39
	$Persist_{RD,24}$	4.18	4.42	4.28	1.00	-1.14	-0.62
	$Const_{0mm/hr}$	3.39	3.11	4.12	6.00	5.83	6.00
	$Const_{10mm/hr}$	2.10	0.35	2.99	0.09	-1.19	0.24
	(operational)						
	Prob. Forecast	-0.14	0.56	0.69	-1.08	-0.74	-0.71

Table 5.4.6.3 Comparison of forecast sources: Standardised Differences for Assessment Measures for Probability Forecasts associated with Evening Update forecasts in the Thames Region. (Rainfall Rates) (In this Table, the base forecast is "Prob. Forecast": the operational probability forecast)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Continuous Brier Score (mm/hr)	*(100% error)*						
	Most Likely	0.14	-0.56	-0.69	1.08	0.74	0.71
	Prob. Median	1.23	-0.72	0.42	2.13	1.33	1.02
	$Persist_{RG,18}$	1.48	-0.11	1.56	-1.00	-0.30	0.73
	$Persist_{RG,24}$	1.71	2.18	2.26	1.01	2.23	0.58
	$Persist_{RD,18}$	4.04	4.43	3.76	-0.66	-1.09	-1.23
	$Persist_{RD,24}$	4.12	4.01	4.14	1.26	-0.91	-0.47
	$Const_{0mm/hr}$	3.02	1.30	3.89	5.60	4.76	5.13
	$Const_{10mm/hr}$	1.88	-0.31	2.72	0.88	-0.39	0.69

5.4.7 Summary

Section 5.4 has described the results obtained for a case study concerning Evening Update forecasts of rainfall in the Environment Agency's Thames Region. The targets of the operational rainfall forecasts in this case are of two types: the largest rainfall accumulation at any site and the largest rainfall rate at any site. These forecasts are given for a single fixed 18-hour time-period and for three Areas of the Thames Region.

The assessments have been performed on a total of 82 forecast-occasions in the case of rainfall accumulations, and 53 occasions in the case of rainfall rates. The forecasts analysed were selected on the basis of lists of prominent rainfall events provided by all the Environment Agency Regions, so that not all the events used would necessarily have contained high rainfall for Thames Region. The different numbers of forecast-occasions for the two types of target arises from an evident confusion of the intention of certain fields within the forecast schema for the early part of the case-study period.

Two sources of ground-truth have been considered in the case of rainfall-rates: a raingauge network and the Nimrod-Quality Controlled radar product. The differences in spatial resolution provided by these products means that the corresponding ground-truths for spatial maximum rainfall amount differ markedly, with the radar source usually providing higher spatial maximum rates than the raingauges. The comparison of the forecasts for rainfall rates against these two sources of ground-truth has shown that the forecasts are much better attuned to the ground-truth provided by raingauges than they are to the radar source. Firstly, the ranges of values of the operational forecasts agree better with the range of values from the raingauges, than the range of radar-derived values. Further, the operational forecasts have a better correlation with the raingauge values than they do with the radar-derived values although the

difference is small. The performance measures which assess the size of forecast-errors are all much better for the raingauge ground-truth than for the radar ground-truth. No firm conclusion can be drawn here because the Nimrod product was still under development during the time-period of these case-study events, particularly in relation to the sets of raingauges available for operational adjustment. However, the explanation that the difference in sizes of the spatial maximum arises from the difference in spatial resolution seems convincing. The important question here is the specification of the forecast target as the spatial maximum: the actual requirement on the Agency's part should first be confirmed. Given that the requirement is for a forecast that will match the values obtained as observations from the Nimrod product, it seems that the Met Office procedures for forecasting spatial maximum should be adjusted against this target. For the purposes of this report, the apparent disparity between the forecasts and outcomes when the radar-based ground-truth is used suggests that the raingauge-based ground-truth should be used for any conclusions.

Results have been presented for a large range of measures of forecast performance, and the results have included comparisons of the operational forecasts against two types of simple forecasts (persistence forecasts and constant-valued forecasts). In the case of forecasts of rainfall amounts, the operational forecasts have been shown to perform better than both of these types of simple forecasts. The operational forecasts of the maximum rainfall rate appear to perform somewhat less well than those for the maximum rainfall totals, at least when the R^2-type of performance statistics are considered. However the comparison is problematic because of the different sets of forecast occasions being considered. The operational forecasts of the maximum rainfall rate again seem to perform better than simple forecasts, although the smaller dataset here than for rainfall totals means that this conclusion is not strongly supported. The persistence forecast for maximum rainfall rate derived from the maximum rate in the immediately preceding 18-hour period turns out to be a fairly strong contender as a forecast.

Besides providing straightforward single-value forecasts, the Evening Updates for this case study provide forecasts in the form of probability tables for the outcome that might occur. The analysis here has included an assessment of these probability forecasts. The results found suggest that these operational probability forecasts do not really perform better than a simple alternative probability forecast derived from the single-value stated as the main forecast (see Section 5.4.1).

The results shown for this case-study have included values for the standardised difference statistic which was described in Section 4.3. This statistic is designed for use in assessing whether there is enough evidence to support a conclusion that one forecast sources is better than another, given that a direct comparison is subject to sampling uncertainty. The utility of this type of statistic has been successfully illustrated.

5.5 Assessment of Heavy Rainfall Warnings

5.5.1 Approach to Assessment

The forecasts provided for the Heavy Rainfall Warning service can be characterised as follows, at least for the warnings issued by the London Weather Centre to Thames, Anglian and Southern Regions. Warnings issued for Thames Region have been chosen for this case study. The Heavy Rainfall Warnings are issued on a irregular basis and cover a single variable-length time-period. Forecasts are provided for two target quantities: the largest rainfall accumulation for the time-period within an area and the highest rainfall intensity within an area over the time-period. In addition, the warning contains a separate indication of a time-period within which the maximum rainfall is expected to be. Besides giving values for the "most likely" outcomes of the maximum amount and maximum rate, the forecasts include brief tables expressing the probabilities that selected threshold values will be exceeded. In both instances, the forecasts relate specifically to spatial maxima rather than to spatial averages. Warnings are issued separately to each of 3 areas sub-dividing each of the 3 Regions that receive Heavy Rainfall Warnings from London Weather Centre. In practice, the warnings are sometimes issued quite a while after the beginning of the nominal forecast-period for which the warning is raised. Here the term "nominal forecast-period" refers to the time-period indicated in the Heavy Rainfall Warning against the caption "Time of event" or "Timing of event", or derived from the values against the captions "Start of Event" and "End of Event", depending on the version of the format being used at the time. The relative timing of issuance has varied from 8 hours before to 2½ hours after the beginning of the nominal interval for the present case study, and the forecast time periods have varied in length from 3 to 21 hours.

The availability in text-file form of the forecast information for Thames Region has again led to this Region being selected for this part of the case study. Although the formats of the files have changed over the various event periods, it has proven possible to adopt an automatic procedure which, in principle, allows the warnings issued during all of the time-periods in Table 5.1.2 to be included in an overall assessment of performance. However, as was the case with the Evening Update forecasts, examination of the Heavy Rainfall Warnings suggested that there had been a problem in interpretation of what was required for the forecasts of rainfall intensities until July 2002 (see Section 5.41). In this instance, since there was only one occasion within the events for analysis when a Warning was issued prior to July 2002, it was decided to omit this one from consideration for assessments of both rainfall amounts and rainfall intensities. Differing numbers of Warnings have been issued for the 3 sub-areas of Thames Region, and the dataset for analysis contains 18 warnings for each of the Northeast and Southeast areas and 13 for Western area.

Ground Truth

For this case study, essentially because it is centred on the same Region of the Environment Agency, the available sources of "ground truth" data are the same as for the assessment of Evening Updates, as discussed in Section 5.4.1. Thus, the network of telemetering raingauges used for operational flood forecasting within the Thames Region provides 47, 28 and 25 raingauges in the Northeast, Southeast and West areas of the Region, respectively. The data were provided as 15 minute accumulations and were processed to form the accumulations and maximum rainfall intensities (over the time-period stated in each warning) for each gauge, from which the spatial maxima were formed. Given this source of ground-truth data, the rainfall intensities derived relate to average intensities over 15 minute time-periods. The second potentially available source of ground-truth is weather radar. Notionally, this might provide a better source of ground-truth data than the raingauge network because of its superior spatial coverage. However, quantitative estimates of rainfall from weather radar are not always reliable. As for the case study for Evening Updates, we have used the radar data to derive values for the maximum rainfall rates within each area, but not for the rainfall accumulations. As before, the radar data are taken from the Nimrod quality controlled "actual" product. The comparisons made here should be treated with caution because this particular radar-product was still under operational development during the time-period used for the assessment. In particular, full sets of quality control procedures may not have been in place, and the availability of raingauge information for adjustment is unclear: either of these two aspects of the Nimrod product may have changed during the assessment period.

The precise definition of the target for the forecast of rainfall intensity is unclear, but discussions with EA staff have indicated that, as for Evening Updates, they interpret these values in relation to what might have been seen in a radar-based rainfall display of rainfall rates. Such display values are based on instantaneous snapshots of rainfall intensity made at either a 5 or 15 minute time-interval. The Nimrod rainfall product is available at a 15-minute time-step, in a form which is a composite of 1, 2 and 5km resolutions. The quantity derived from the Nimrod product for comparison against the rainfall intensity component of the Heavy Rainfall Warning was the maximum of all the 15-minute rainfall values falling within the forecast period and within the particular sub-area of the Thames Region (for 1 km pixels entirely within the sub-area). Since the Warnings contain an indication of when the maximum rainfall rate is expected, in the form of a time-period which is usually strictly contained within the overall time-period quoted for the warning, it is possible to compare the maximum rainfall rate quoted in the forecast against observed maximum rates for either the overall period or for the period specifically indicated to contain the maximum rate.

Comparative Forecasts

The content of the Heavy Rainfall Warnings issued by the London Weather Centre is very similar to that of the routine Evening Updates, and a similar approach has been taken to that reported in Section 5.4.1. Nominally the Heavy Rainfall Warnings provide two separate forecasts for both maximum rainfall accumulation and maximum rainfall rate, one of which is an ordinary, single-valued forecast and the other a probability forecast. However the probability forecast has been used to derive a second single-valued forecast from the probability distribution: the median of this

distribution is used. As before, the method used to define the median was based on linear interpolation in the probability table, rather than being based on fitting some parametric distribution. While the median is not a serious contender as an operational forecast, it does provide an example for this study of a forecast which should be close in performance to the "most likely" value contained in the Heavy Rainfall Warnings and which can therefore be used to illustrate the difficulties involved in performing a careful comparison of closely matched forecast sources.

The majority of the forecast-periods for this case study are longer than 6 hours and hence the networked radar products do not provide a source of comparative forecasts. While the mesoscale model may eventually provide a possible alternative source of forecasts, data from this source were not available for the present phase of the study.

To summarise, for this phase of the project, the main set of forecasts that are available for comparison are all derived from the Heavy Rainfall Warnings and are:

(i) the explicit forecast indicated as "most likely value";
(ii) a derived forecast, calculated as the median of the probability forecast;
(iii) the probability forecast itself.

Thus there are two single-valued forecasts and one probability forecast. The situation here is the same as for the Evening Update forecasts discussed in Section 5.4.1. Once again the set of candidates has been extended in two ways. Firstly, by defining some additional single-valued forecasts of a rather simple nature and, secondly, by defining some additional probability forecasts which can be derived from the single-valued forecasts in a simple way. It is convenient to treat the assessment of single-valued forecasts and probability forecasts as separate tasks, but it should be noted that among the simple probability forecasts are some which correspond to expressing absolute certainty about a single value.

Two types of simple forecasts have been included for comparison. In the first of these, a constant value is used directly as the forecast, while for the second the value used as the forecast is constructed to be proportional to the interval length. The constants defining the forecasts have been set to give forecasts of about the same size as the forecasts contained in the Heavy Rainfall Warnings, but there has been no attempt to tune these values to give good performance.

As for the Evening Updates discussed in Section 5.4.1, the simple probability forecasts included for comparison are of two types. For the first type, the single-valued forecasts outlined above are included with the probability component of the forecast constructed so as to express absolute certainty in the single-value forecast. The second type of probability forecast is again constructed from the single-valued forecast, but with the uncertainty in the forecast being determined by the rule that the probability is uniformly distributed over an interval centred on the single-valued forecast with a width that is the same as the central value (i.e. from 0 to 200% of the central value), with an overriding minimum of 1 unit (mm or mm h^{-1}, depending on the quantity being forecasted). For instances where this interval extends to negative values, the probability distribution is revised so that the probability for negative values is replaced by a discrete component of probability for the value zero. The choice of the size of the interval used here is entirely arbitrary and there may be better ways of associating a probability with the single-valued forecasts.

Table 5.5.1.1 provides a summary of the ground-truth and comparative forecasts that are available for this study for the rainfall-accumulation component of the Evening Update forecast. Table 5.5.1.2 provides a similar summary for the maximum rainfall rate component of the Heavy Rainfall Warnings. It should be noted that the forecasts labelled "$Const_{2mm/hr}$" and "$Const_{4mm/hr/hr}$" are not constant-valued forecasts: rather, the forecast-values are constructed to be proportional to the time-interval length. For example, if the nominal forecast period had an interval length of 10 hours, then the value use for the forecast of maximum rainfall accumulation would be 20mm (derived as 2 mm h^{-1} times 10 hours), while the forecast of maximum rainfall rate would be 40 mm h^{-1} (derived as 4 mm h^{-2} times 10 hours).

Table 5.5.1.1 Summary of Assessment for Heavy Rainfall Warning forecasts of Maximum Rainfall Accumulations

Description	Abbreviation
Ground truth	
Maximum accumulation across raingauges in area	
Single-valued forecasts	
Operational candidates	
Values labelled 'most likely' in Heavy Rainfall Warning	Most Likely
Median of probability forecast in Heavy Rainfall Warning	Prob. Median
Comparative forecasts	
A value for the maximum accumulation constructed as twice the time-period length in hours.	$Const_{2mm/hr}$
A fixed value of 20 mm for the maximum accumulation	$Const_{20mm}$
Probability forecasts	
Operational candidates	
Probability Forecast from Heavy Rainfall Warning	Prob. Forecast
Comparative forecasts	
The single-valued forecasts listed above treated as being absolutely certain	(certain)
The single-valued forecasts listed above, with uncertainty uniform over ± 100% or ± 1mm, whichever is larger.	(100% error)

Table 5.5.1.2 Summary of Assessment for Heavy Rainfall Warning forecasts of Maximum Rainfall Rates

Description	Abbreviation
Ground truth	
Maximum of all 15-minute accumulations at raingauges in the area in the time-period, converted to rate	
Maximum 15-minute rainfall rate in the time-period in the area as estimated by the Nimrod radar product	
Single-valued forecasts	
Operational candidates	
'Most likely' from Heavy Rainfall Warning	Most Likely
Median of probability forecast in Heavy Rainfall Warning	Prob. Median
Comparative forecasts	
A value for the maximum rate constructed as four times the time-period length in hours.	$Const_{4mm/hr/hr}$
A fixed value of 30 mm h^{-1} for the maximum rate	$Const_{30mm/hr}$
Probability forecasts	
Operational candidates	
Probability Forecast from Heavy Rainfall Warning	Prob. Forecast
Comparative forecasts	
The single-valued forecasts listed above treated as being absolutely certain	(certain)
The single-valued forecasts listed under (i) to (viii) with uncertainty uniform over ± 100% or ± 1mm h^{-1}, whichever is larger.	(100% error)

5.5.2 Example Forecasts and Outcomes

Table 5.5.2.1 lists the full set of data for the assessment of forecasts for the Northeast area of the Agency's Thames Region in the case of the maximum rainfall accumulation forecast. The dates and times reported here indicate the start of the forecast period. Times have been converted to GMT.

Table 5.5.2.1 Example of data for assessment of rainfall forecasts for rainfall accumulations: maximum totals in Northeast area of Thames Region (units: mm).

date issued	period start	period end	most likely	median	2mm /hr	Outcome from Raingauges
29 7 2002 16:30	29 7 2002 16:00	29 7 2002 23:00	15.00	10.00	14.00	3.60
30 7 2002 20:36	30 7 2002 20:30	31 7 2002 05:00	25.00	25.00	17.00	28.20
31 7 2002 11:20	31 7 2002 11:00	31 7 2002 23:00	15.00	10.00	24.00	26.60
3 8 2002 13:24	3 8 2002 15:00	3 8 2002 20:00	10.00	7.50	10.00	22.00
3 8 2002 22:28	3 8 2002 20:00	4 8 2002 02:00	20.00	10.00	12.00	19.80
4 8 2002 11:44	4 8 2002 12:00	4 8 2002 20:00	25.00	25.00	16.00	22.60
5 8 2002 09:42	5 8 2002 10:00	5 8 2002 21:00	30.00	20.00	22.00	11.20
7 8 2002 10:52	7 8 2002 14:00	8 8 2002 08:00	30.00	12.50	36.00	23.40
9 8 2002 04:44	9 8 2002 05:00	9 8 2002 23:00	25.00	15.00	36.00	41.80
10 8 2002 10:52	10 8 2002 11:00	10 8 2002 20:00	25.00	20.00	18.00	6.00
9 9 2002 08:01	9 9 2002 08:00	9 9 2002 17:00	35.00	26.67	18.00	44.20
13 10 2002 06:10	13 10 2002 10:00	14 10 2002 02:00	18.00	18.00	32.00	14.80
15 10 2002 13:30	15 10 2002 13:00	15 10 2002 20:00	15.00	8.33	14.00	22.20
21 10 2002 12:14	21 10 2002 13:00	21 10 2002 16:00	15.00	10.00	6.00	7.80
22 10 2002 06:27	22 10 2002 07:00	22 10 2002 17:00	20.00	11.67	20.00	14.80
26 10 2002 10:42	26 10 2002 19:00	27 10 2002 03:00	12.00	10.00	16.00	4.80
2 11 2002 07:26	2 11 2002 12:00	2 11 2002 23:00	15.00	16.67	22.00	15.60

The values given in Table 5.5.2.1 can be used to compare the two single-valued forecasts derived from the Heavy Rainfall Warnings: the 'most likely' value, quoted directly in the forecast, and the median of the probability forecast. These values do tend to vary together in a reasonable way, but there are often sizeable differences and these are larger than found for the Evening Updates in Table 5.4.2.1.

Figures 5.5.2.1 to 5.5.2.3 provide a complete set of scatter plots of forecast-values against outcomes for the present case study and for the case of forecasts of rainfall amounts. Once again, these plots indicate that there is not a particularly good correspondence between the operational forecasts and the outcomes as derived from the raingauge network. More importantly for the purposes of the analysis here, it is not the case that the performance analyses will be completely dominated by only one or two particularly bad forecasts. There is some interest in the behaviour of the simple forecast derived by setting the forecast-value to be proportional to the interval length: the plots here show that there is no clear distinction between this simple forecast and the operational forecasts.

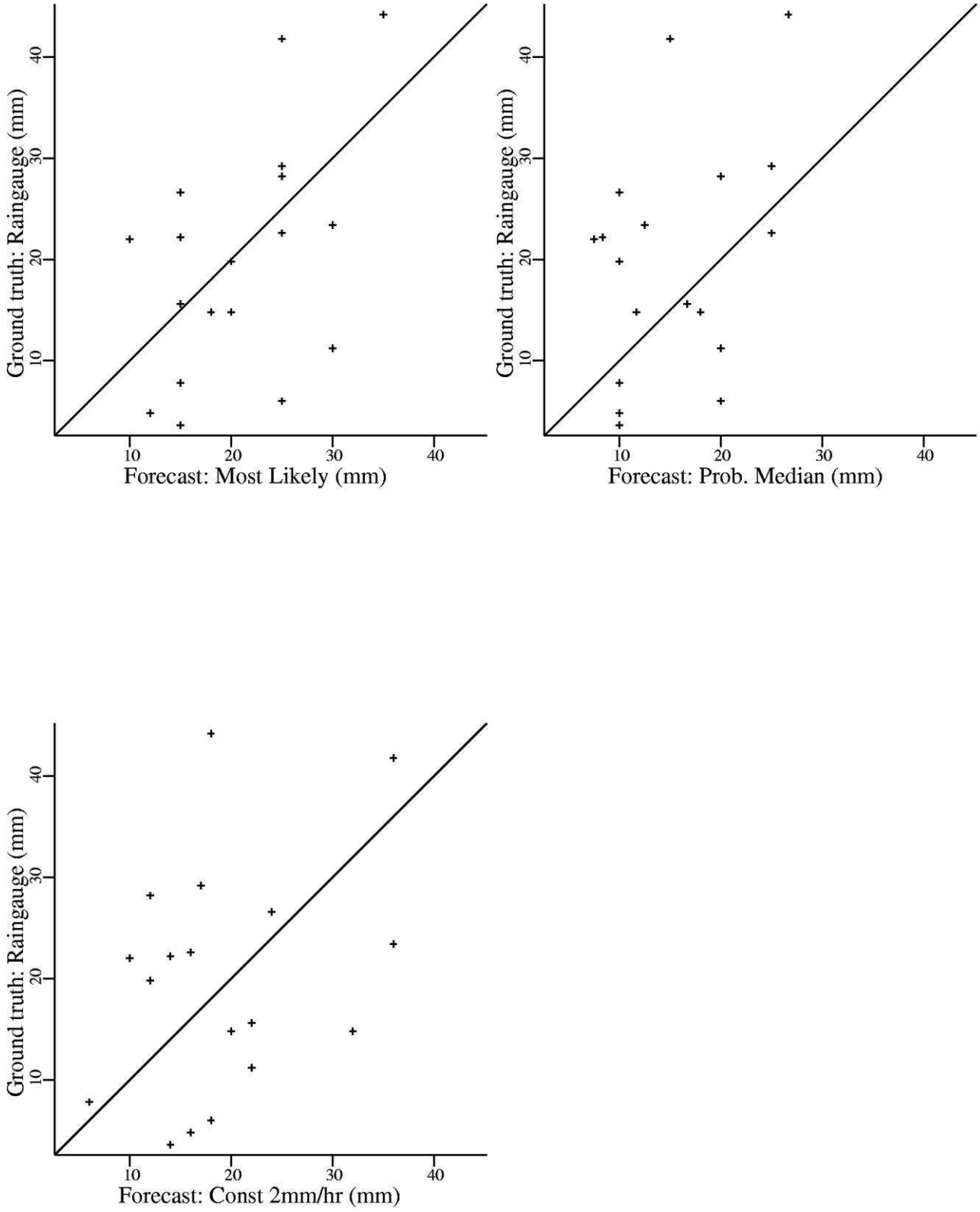


Figure 5.5.2.1 Heavy Rainfall warning forecasts of maximum rainfall amounts. Ground truth from raingauge network. Northeast sub-area of Thames Region

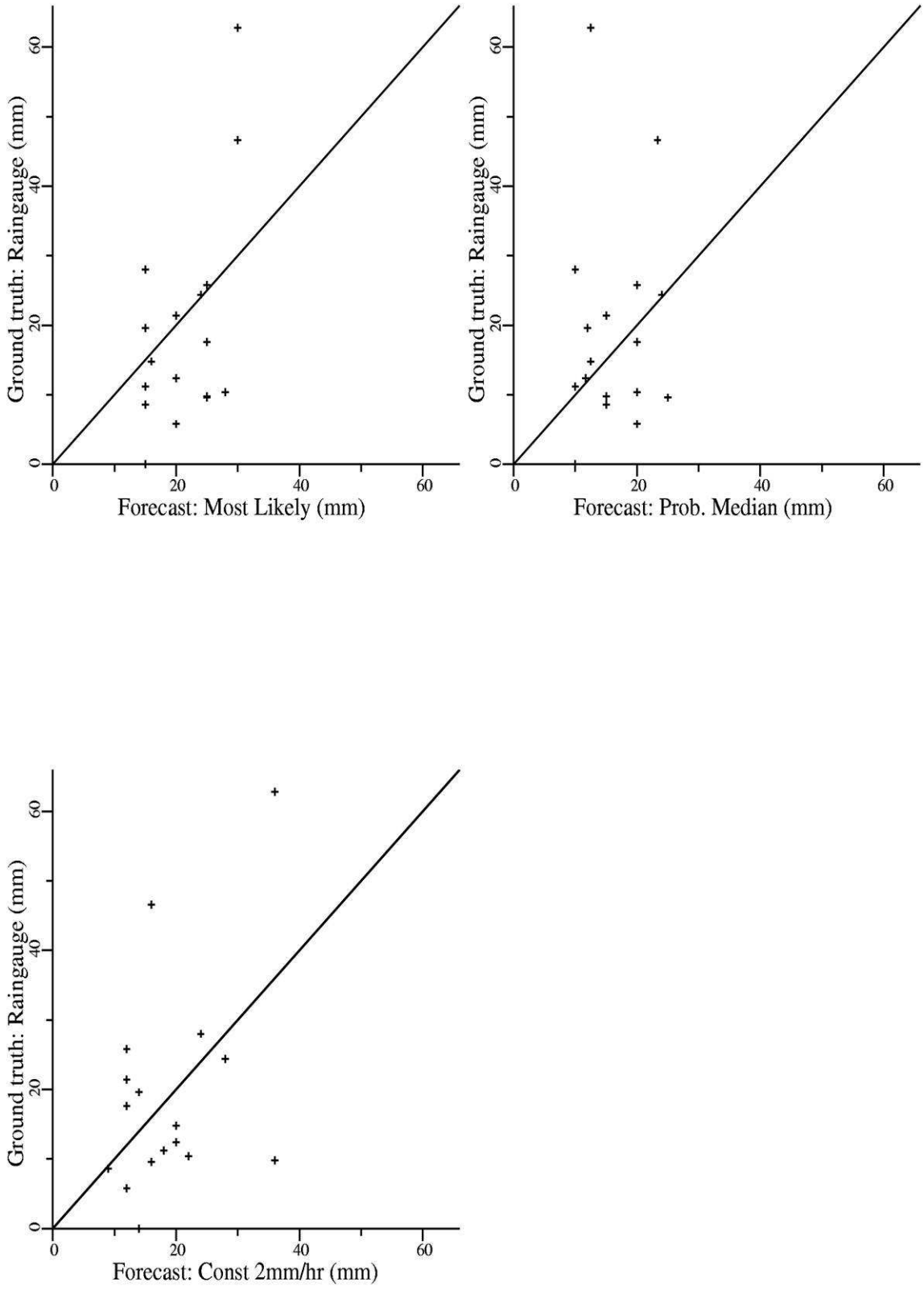


Figure 5.5.2.2 Heavy Rainfall warning forecasts of maximum rainfall amounts. Ground truth from raingauge network. Southeast sub-area of Thames Region

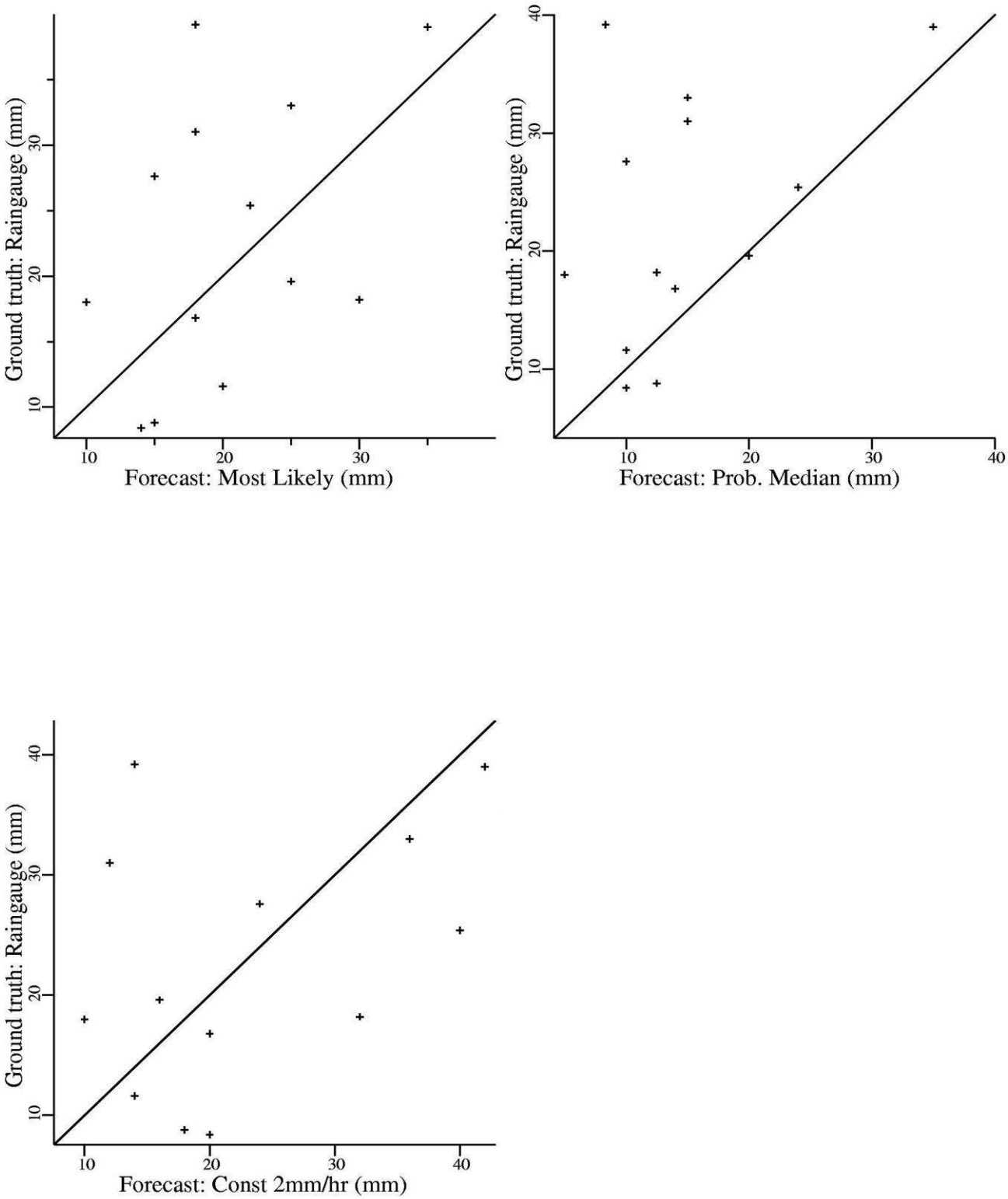


Figure 5.5.2.3 Heavy Rainfall Warning forecasts of maximum rainfall amounts. Ground truth from raingauge network. Western sub-area of Thames Region

Table 5.5.2.2 Example of data for assessment of rainfall forecasts for maximum rainfall rates in the overall forecast time-period: maximum rate in Northeast area of Thames Region (units: mm h^{-1}).

date issued	period start	period end	most likely	median	4mm /hr/hr	--Outcome from-- R'gauge	Radar
29 7 2002 16:30	29 7 2002 16:00	29 7 2002 23:00	20.00	26.00	28.00	5.60	191.75
30 7 2002 20:36	30 7 2002 20:30	31 7 2002 05:00	15.00	27.50	34.00	49.60	76.09
31 7 2002 05:16	31 7 2002 05:00	31 7 2002 11:00	20.00	20.00	24.00	27.20	37.28
31 7 2002 11:20	31 7 2002 11:00	31 7 2002 23:00	25.00	12.50	48.00	77.60	79.12
3 8 2002 13:24	3 8 2002 15:00	3 8 2002 20:00	12.00	8.50	20.00	52.80	109.56
3 8 2002 22:28	3 8 2002 20:00	4 8 2002 02:00	8.00	8.00	24.00	30.40	66.97
4 8 2002 11:44	4 8 2002 12:00	4 8 2002 20:00	32.00	30.00	32.00	45.60	124.78
5 8 2002 09:42	5 8 2002 10:00	5 8 2002 21:00	32.00	42.50	44.00	26.40	76.09
7 8 2002 10:52	7 8 2002 14:00	8 8 2002 08:00	15.00	8.00	72.00	42.40	133.94
9 8 2002 04:44	9 8 2002 05:00	9 8 2002 23:00	24.00	20.00	72.00	56.80	170.44
10 8 2002 10:52	10 8 2002 11:00	10 8 2002 20:00	25.00	20.00	36.00	19.20	88.28
9 9 2002 08:01	9 9 2002 08:00	9 9 2002 17:00	30.00	30.00	36.00	41.60	68.56
13 10 2002 06:10	13 10 2002 10:00	14 10 2002 02:00	32.00	35.00	64.00	7.20	19.31
15 10 2002 13:30	15 10 2002 13:00	15 10 2002 20:00	10.00	14.00	28.00	25.60	28.19
21 10 2002 12:14	21 10 2002 13:00	21 10 2002 16:00	12.00	8.50	12.00	30.40	26.19
22 10 2002 06:27	22 10 2002 07:00	22 10 2002 17:00	15.00	13.08	40.00	17.60	36.41
26 10 2002 10:42	26 10 2002 19:00	27 10 2002 03:00	6.00	6.00	32.00	4.00	12.56
2 11 2002 07:26	2 11 2002 12:00	2 11 2002 23:00	8.00	12.50	44.00	6.40	62.91

Table 5.5.2.3 Example of data for assessment of rainfall forecasts for maximum rainfall rates in the time-period forecasted to contain the maximum rate: Northeast area of Thames Region (units: mm h^{-1}).

date issued	period start	period end	most likely	median	4mm /hr/hr	--Outcome from-- R'gauge	Radar
29 7 2002 16:30	29 7 2002 16:00	29 7 2002 21:00	20.00	26.00	20.00	5.60	191.75
30 7 2002 20:36	30 7 2002 20:30	30 7 2002 23:00	15.00	27.50	10.00	49.60	76.09
31 7 2002 05:16	31 7 2002 05:00	31 7 2002 11:00	20.00	20.00	24.00	27.20	37.28
31 7 2002 11:20	31 7 2002 13:00	31 7 2002 19:00	25.00	12.50	24.00	77.60	79.12
3 8 2002 13:24	3 8 2002 15:00	3 8 2002 20:00	12.00	8.50	20.00	52.80	109.56
3 8 2002 22:28	3 8 2002 22:00	4 8 2002 01:00	8.00	8.00	12.00	20.00*	35.00*
4 8 2002 11:44	4 8 2002 12:00	4 8 2002 17:00	32.00	30.00	20.00	45.60	124.78
5 8 2002 09:42	5 8 2002 12:00	5 8 2002 18:00	32.00	42.50	24.00	26.40	73.06*
7 8 2002 10:52	7 8 2002 14:00	8 8 2002 02:00	15.00	8.00	48.00	42.40	133.94
9 8 2002 04:44	9 8 2002 14:00	9 8 2002 21:00	24.00	20.00	28.00	56.80	170.44
10 8 2002 10:52	10 8 2002 11:00	10 8 2002 18:00	25.00	20.00	28.00	19.20	88.28
9 9 2002 08:01	9 9 2002 10:00	9 9 2002 13:00	30.00	30.00	12.00	41.60	29.88*
13 10 2002 06:10	13 10 2002 11:00	13 10 2002 14:00	32.00	35.00	12.00	7.20	6.09*
15 10 2002 13:30	15 10 2002 13:00	15 10 2002 17:00	10.00	14.00	16.00	25.60	28.19
21 10 2002 12:14	21 10 2002 13:00	21 10 2002 15:00	12.00	8.50	8.00	30.40	26.19
22 10 2002 06:27	22 10 2002 07:00	22 10 2002 10:00	15.00	13.08	12.00	11.20*	13.44*
26 10 2002 10:42	26 10 2002 21:00	27 10 2002 02:00	6.00	6.00	20.00	4.00	12.56
2 11 2002 07:26	2 11 2002 18:00	2 11 2002 23:00	8.00	12.50	20.00	6.40	62.91

* occasions when the outcome for the time-period forecasted to contain the maximum value differs from that for the overall forecast-period.

Table 5.5.2.2 lists the full set of data for the assessment of forecasts for the North East area of the Agency's Thames Region in the case of the maximum rainfall rates in the overall forecast-period. As for the Evening Updates (Section 5.4.1), it will be seen that values for the spatial maximum of rainfall rates obtained from weather radar are usually substantially higher than those obtained from the network of raingauges. Table 5.2.3 provides a similar listing, but in this case the time-period used is that quoted in the Warnings as likely to contain the maximum rainfall rate. These results

show that, when the raingauge network is used to provide the ground-truth, the time-periods that are forecasted to contain the maximum rainfall rates are usually quite successful. No further analysis has been made of the performance of the forecasts of rainfall rates when the outcome is judged on the interval that is forecasted to contain the maximum rate: Table 5.2.3 indicates that there would be little change from the results when the full forecast-period is used.

The complete set of scatter plots for the case of Heavy Rainfall Warning forecasts of maximum rainfall rates, is provided in Figures 5.5.2.4 to 5.5.2.9. Once again, these plots indicate that there is not a particularly good correspondence between the operational forecasts and the eventual outcomes and that the performance analyses will not be completely dominated by only one or two particularly bad forecasts. The correspondence between the operational forecasts and the radar-derived ground truth is seen to be particularly poor, with the forecast-values never extending even into the mid-range of the values of the outcomes derived from radar. As for the case of forecasts of maximum rainfall amount (Figures 5.5.2.1 to 5.5.2.3), there is some interest in the behaviour of the simple forecast derived by setting the forecast-value to be proportional to the interval length: again the plots here show that there is no clear distinction between this simple forecast and the operational forecasts.

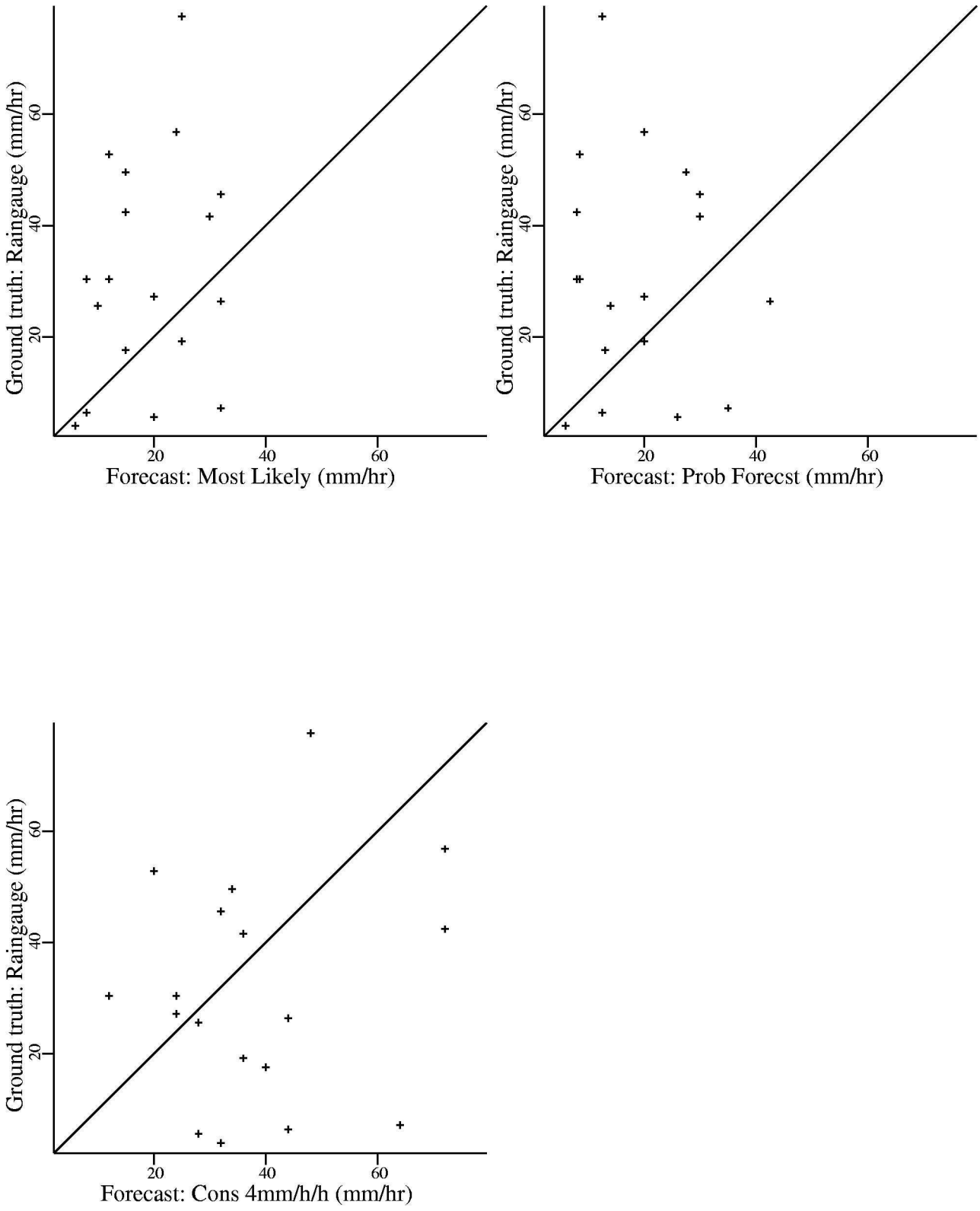


Figure 5.5.2.4 Heavy Rainfall warning forecasts of maximum rainfall rates. Ground truth from raingauge network. Northeast sub-area of Thames Region

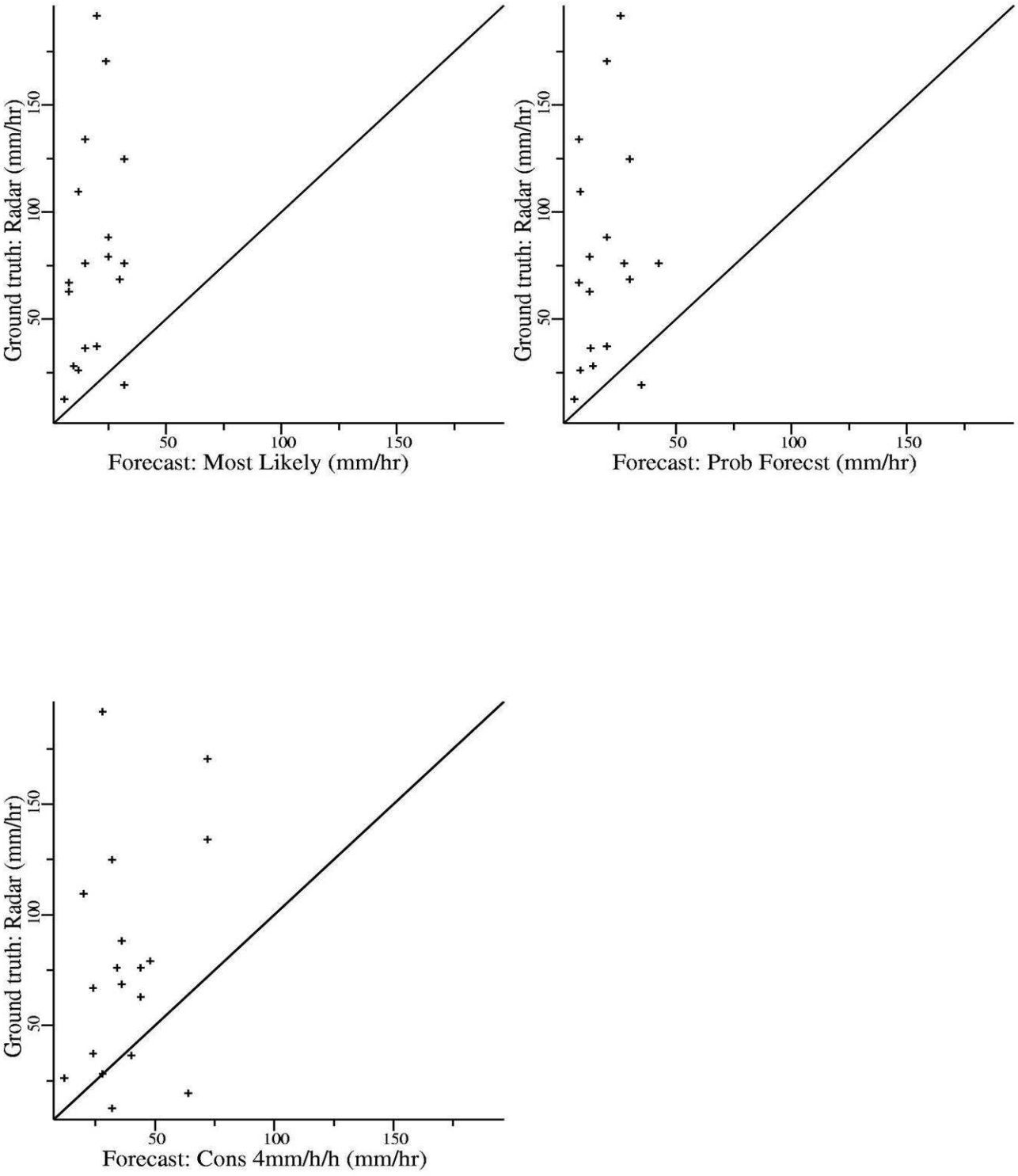


Figure 5.5.2.5 Heavy Rainfall warning forecasts of maximum rainfall rates. Ground truth from Nimrod QC radar. Northeast sub-area of Thames Region

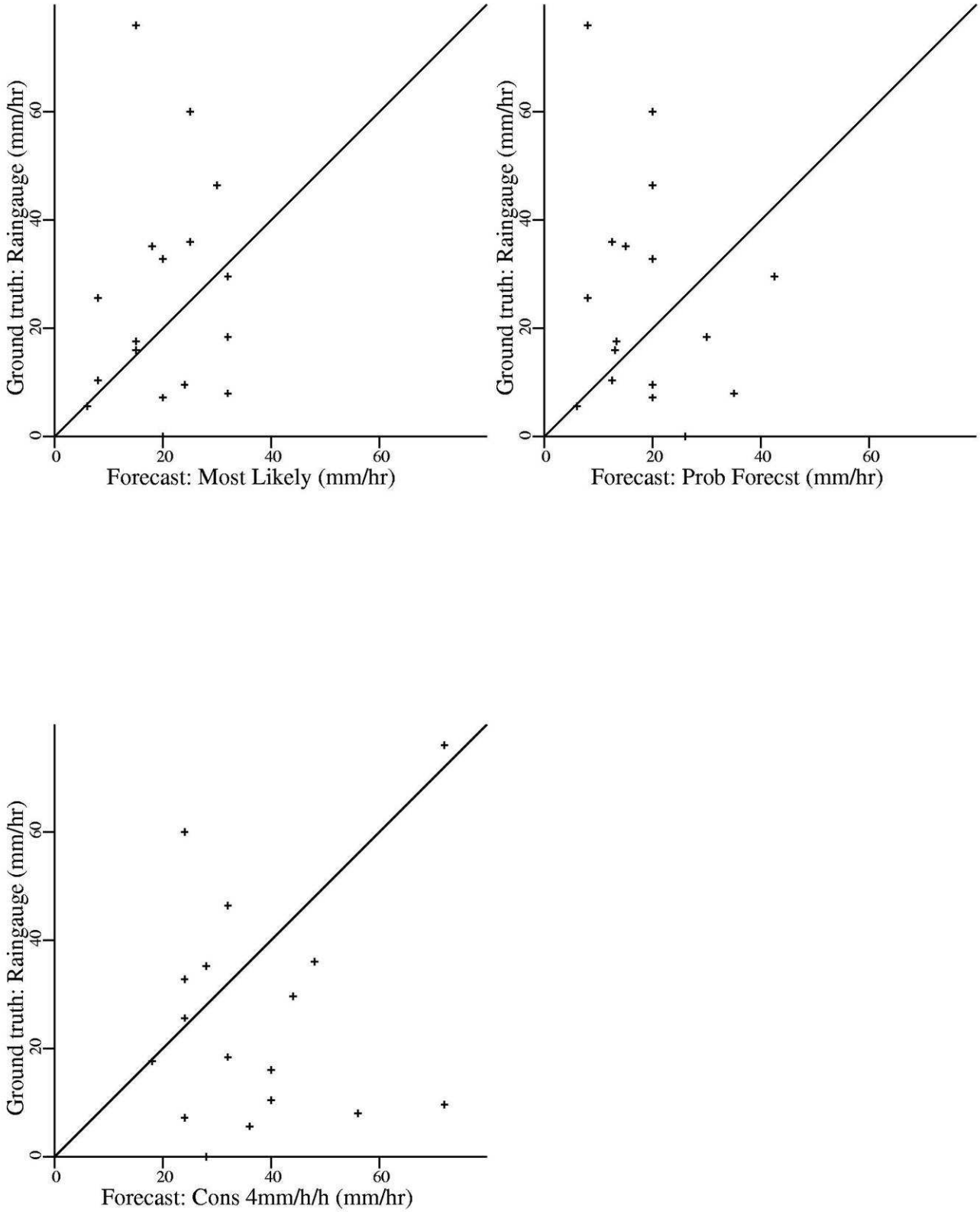
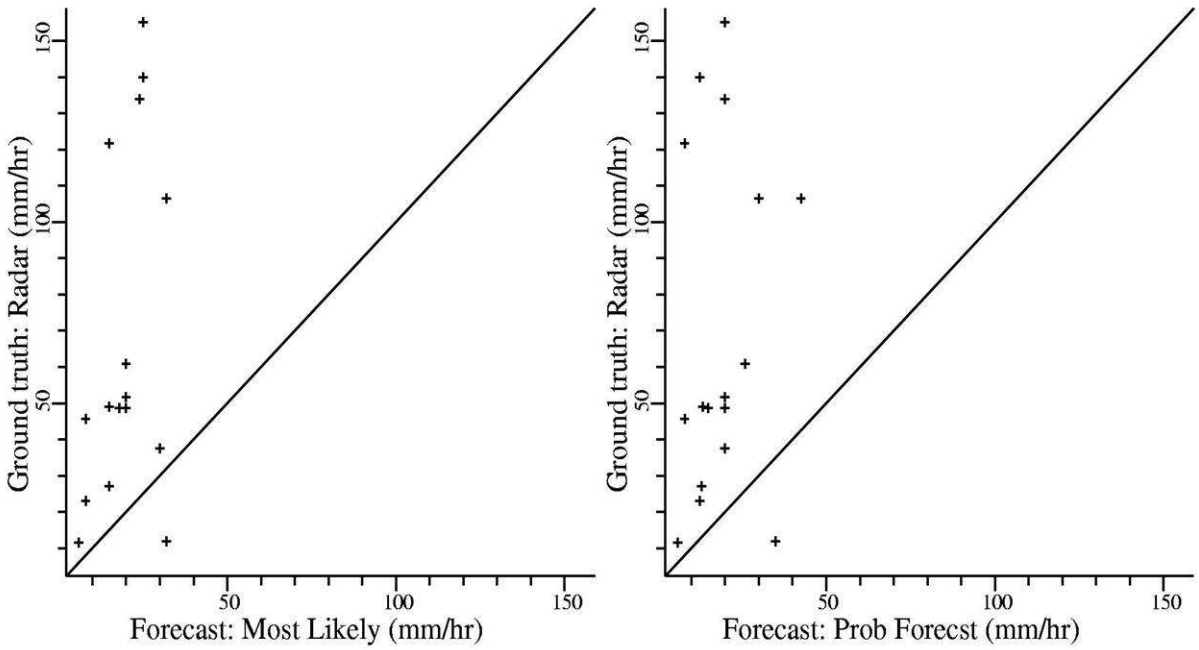


Figure 5.5.2.6 Heavy Rainfall warning forecasts of maximum rainfall rates. Ground truth from raingauge network. Southeast sub-area of Thames Region



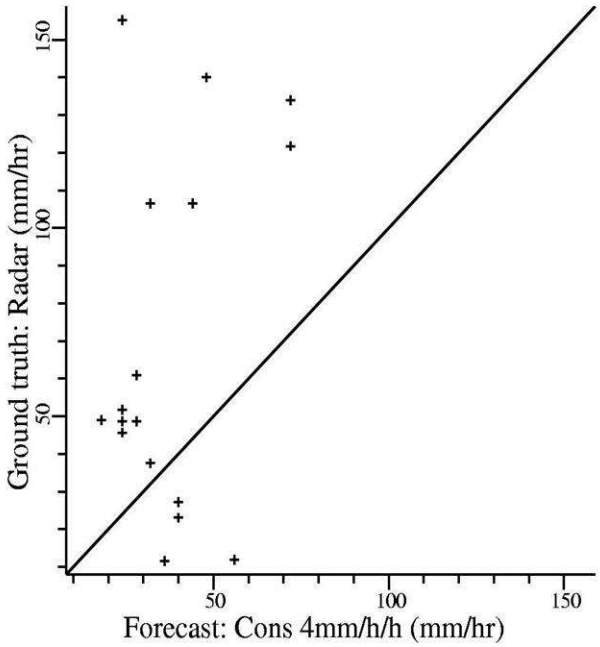


Figure 5.5.2.7 Heavy Rainfall warning forecasts of maximum rainfall rates. Ground truth from Nimrod QC radar. Southeast sub-area of Thames Region

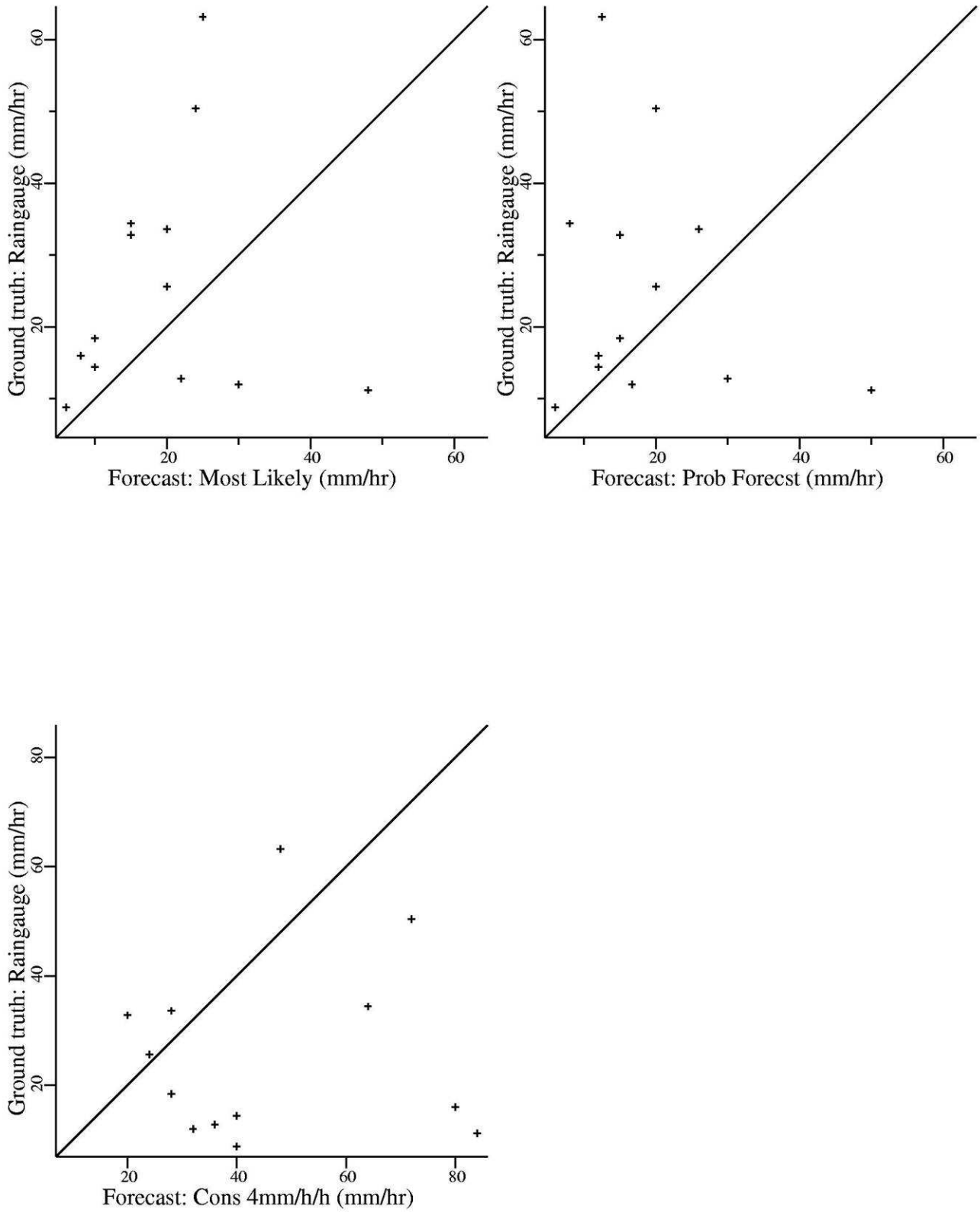


Figure 5.5.2.8 Heavy Rainfall warning forecasts of maximum rainfall rates. Ground truth from raingauge network. Western sub-area of Thames Region

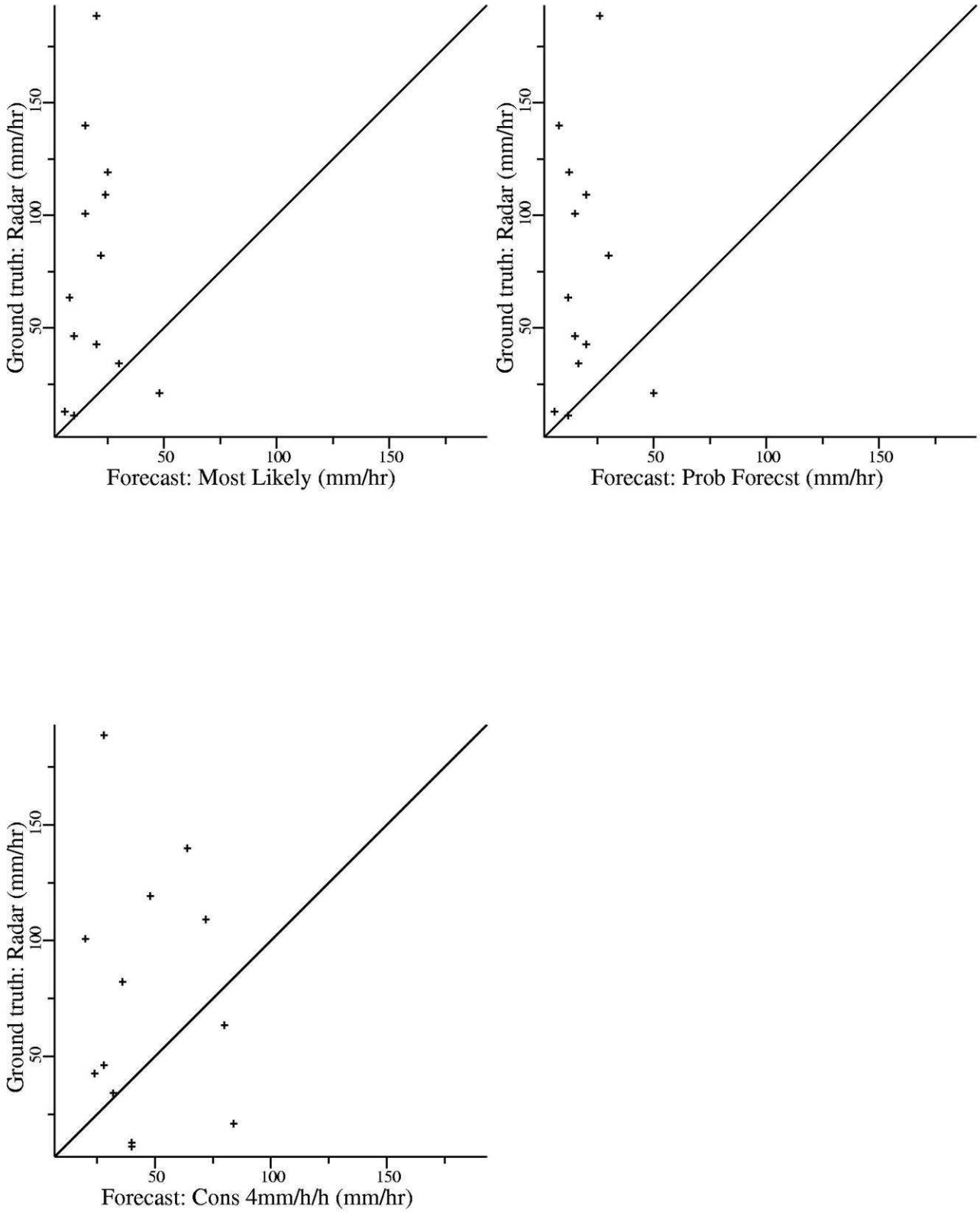


Figure 5.5.2.9 Heavy Rainfall warning forecasts of maximum rainfall rates. Ground truth from Nimrod QC radar. Western sub-area of Thames Region

5.5.3 Assessment of Single-valued Forecasts of Accumulations

5.5.3.1 Assessment of forecast amounts

Section 2.2.3 has outlined a number of measures of forecast performance appropriate for single-valued forecasts of rainfall amounts. Several of these have been evaluated for the Heavy Rainfall Warning forecasts for Thames Region, and the results are presented in Tables 5.5.3.1.1-6. The set of performance measures is the same as that used for the Evening Update forecasts which were discussed in Section 5.4.3.

Table 5.5.3.1.1 shows the basic assessment measures for the size of forecast errors for rainfall amounts, evaluated for the 3 sub-areas of the Thames Region. Results are given for the 4 types of single-valued forecasts listed in Table 5.5.1.1 and, in addition, the result is given for the best performance measure obtainable by a constant-value forecast (rows labelled "$Const_{best}$"). Table 5.5.3.1.2 shows the corresponding R^2 (efficiency) measures: these effectively compare the values of the performance measures shown in Table 5.5.3.1.1 with the best performance measure achievable by a constant-value forecast.

The results in Tables 5.5.3.1.1 and 5.5.3.1.2 illustrate that the "Most Likely" forecast contained in the Heavy Rainfall Warnings provided the best forecast performance according to the majority of the performance measures, across the 3 sub-areas being investigated. Performance of the Median value extracted from the probability forecasts is quite a lot worse than that of the "Most Likely" forecast: this contrasts with the result found for the Evening Updates where the corresponding median forecasts were only a little worse than the "Most Likely" forecast. The R^2 (efficiency) measures for the operational candidate forecasts are small and sometimes negative, which indicates that the forecasts are not really doing a lot better (if at all) than could be achieved by using a suitable constant-value as the forecast of the rainfall amount. Thus, once a forecaster has decided to issue a Warning covering a particular time-period, there is little extra forecasting skill in the estimate of areal-maximum rainfall amount for that time period.

Table 5.5.3.1.3 shows details of the bias contained in the various forecast sources. Here the usual statistical practice is followed of defining the direction in which an "error" is measured as being positive if the outcome is larger than the forecast, and hence the bias being negative means that the forecast tends to be larger than the actual outcome. It can be seen from this table that the Median of the probability forecast tend to be rather smaller than the "Most Likely" forecast by some 5 or 6 mm on average. For the example region given in Table 5.5.2.1, none of the "Prob.Median" forecasts are larger than the "Most Likely" forecast. Table 5.5.3.1.4 shows some simple statistics which give more details of the typical amounts obtained for the actual outcomes and for the forecasts of rainfall amounts. This table again shows that the median forecasts are typically lower than the "Most Likely" forecasts and it indicates that, for most sub-areas, the typical amounts given by the "most likely" forecasts are in better agreement with the typical amounts actually accruing when Heavy Rainfalls Warnings are issued than are the median forecasts. The forecast values have standard-

deviations rather lower than the actual outcomes, a feature which would be expected in most forecasting situations.

Table 5.5.3.1.5 gives values for correlation and regression coefficients for linear relationships between outcomes and forecasts of rainfall and log-rainfall. The interpretation of the coefficients here is similar to that outlined in the discussion of Evening Update forecasts in Section 5.4.3.1. The present case-study has included the "$Const_{2mm/hr}$" forecast in which the forecast is proportional to the interval-length. Table 5.5.3.1.4 shows that this type of forecast can have a correlation with the eventual outcome that is nearly as large as that for the operational forecasts: this seems to indicate that at least part of the skill of the operational forecasts of rainfall amounts arises from getting the event-length correct.

The above analysis of performance has been the traditional one where standard measures of forecast performance are evaluated separately for each forecast source and then compared. As discussed in Section 4.3, it is possible to do a rather more detailed analysis and to determine whether the evidence provided by the test dataset is sufficient to distinguish between the performance of different forecast sources, bearing in mind the sampling variability of the forecast performance statistics and the statistical dependences between them. Table 5.5.3.1.6 relates directly to this question. Taking the 'most likely' forecast ("Most Likely") from the Heavy Rainfall Warnings as a "base forecast", Table 5.5.3.1.6 considers each of the other forecast sources in turn and asks how much evidence there is that the "base forecast" is better than the candidate forecast. The values given are the standardised differences discussed earlier in Section 4.3.3, and positive values indicate that the "base forecast" has a smaller size of error, as measured by the performance statistic, than the candidate. If the alternative candidate forecast produces smaller errors, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the long-run performance measures for the two forecast sources will turn out to be in the order indicated. For the purposes here, a standardised difference outside the range ± 2 units indicates fairly strong evidence that one forecast source really is better than another (because of the small number of forecast-occasions being used here, this might be better replaced by ± 2.1 units, but this has little effect).

The results in Table 5.5.3.1.6 reflect those in Table 5.5.3.1.1, in that the comparisons which favour one forecast source over another are the same. However, Table 5.5.3.1.6 provides extra information. The situation here is the same as in Section 5.4.3.1 where Evening Update forecasts were analysed, except that here there are substantially fewer forecast-occasions on which to base the analysis. It is therefore not surprising that no clear conclusion can be drawn from the comparison of forecast performance.

Table 5.5.3.1.1 Raw assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Totals)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error (mm)	Most Likely	8.11	10.06	8.37
	Prob. Median	9.38	10.00	8.97
	$Const_{2mm/hr}$	10.29	10.63	9.25
	$Const_{20mm}$	8.99	10.93	8.75
	$Const_{best}$	8.99	10.04	8.72
Root Mean Square Error (mm)	Most Likely	9.89	12.86	9.78
	Prob. Median	11.49	15.35	12.68
	$Const_{2mm/hr}$	11.72	13.56	11.58
	$Const_{20mm}$	11.25	14.76	10.58
	$Const_{best}$	11.25	14.72	10.20
Mean Absolute Error of Log-Rainfall (dimensionless)	Most Likely	0.50	0.75	0.42
	Prob. Median	0.60	0.75	0.52
	$Const_{2mm/hr}$	0.62	0.78	0.46
	$Const_{20mm}$	0.54	0.80	0.42
	$Const_{best}$	0.54	0.75	0.42
Root Mean Square Error of Log-Rainfall (dimensionless)	Most Likely	0.67	1.32	0.47
	Prob. Median	0.70	1.27	0.71
	$Const_{2mm/hr}$	0.71	1.31	0.57
	$Const_{20mm}$	0.73	1.39	0.50
	$Const_{best}$	0.70	1.31	0.50

Table 5.5.3.1.2 R^2 (efficiency) measures for Heavy Rainfall Warning forecasts in the Thames Region for each type of assessment measure. (Rainfall Totals)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
R^2 for Mean Absolute Error	Most Likely	0.10	0.00	0.04
	Prob. Median	-0.04	0.00	-0.03
	$Const_{2mm/hr}$	-0.14	-0.06	-0.06
	$Const_{20mm}$	0.00	-0.09	0.00
	$Const_{best}$	0.00	0.00	0.00
R^2 for Root Mean Square Error	Most Likely	0.23	0.24	0.08
	Prob. Median	-0.04	-0.09	-0.55
	$Const_{2mm/hr}$	-0.09	0.15	-0.29
	$Const_{20mm}$	0.00	-0.01	-0.08
	$Const_{best}$	0.00	0.00	0.00
R^2 for Mean Absolute Error of Log-Rainfall	Most Likely	0.07	-0.01	0.00
	Prob. Median	-0.10	0.00	-0.24
	$Const_{2mm/hr}$	-0.14	-0.04	-0.11
	$Const_{20mm}$	0.00	-0.07	0.00
	$Const_{best}$	0.00	0.00	0.00
R^2 for Root Mean Square Error of Log-Rainfall	Most Likely	0.08	-0.03	0.12
	Prob. Median	-0.03	0.05	-1.01
	$Const_{2mm/hr}$	-0.04	0.00	-0.29
	$Const_{20mm}$	-0.09	-0.13	0.00
	$Const_{best}$	0.00	0.00	0.00

Table 5.5.3.1.3 Bias measures for Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Totals)

Bias Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Error (mm)	Most Likely	-0.91	-2.32	2.43
	Prob. Median	4.57	2.97	8.10
	$Const_{2mm/hr}$	0.76	0.01	-0.11
	$Const_{20mm}$	-0.08	-1.04	2.82
Median Error (mm)	Most Likely	-1.30	-5.10	3.42
	Prob. Median	3.67	0.73	4.00
	$Const_{2mm/hr}$	2.20	-4.40	-3.00
	$Const_{20mm}$	0.90	-6.40	-0.40
Mean Error of Log-Rainfall (dimensionless)	Most Likely	-0.19	-0.51	0.05
	Prob. Median	0.14	-0.21	0.44
	$Const_{2mm/hr}$	-0.07	-0.35	-0.02
	$Const_{20mm}$	-0.21	-0.48	0.02
Median Error of Log-Rainfall (dimensionless)	Most Likely	-0.06	-0.32	0.11
	Prob. Median	0.20	0.06	0.18
	$Const_{2mm/hr}$	0.13	-0.22	-0.09
	$Const_{20mm}$	0.04	-0.39	-0.02

Table 5.5.3.1.4 Statistics of forecasts and outcomes for Heavy Rainfall Warning forecasts in Thames Region. (Rainfall Totals)

Statistic of Rainfall	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Rainfall (mm)	Outcome	19.92	18.96	22.82
	Most Likely	20.83	21.28	20.38
	Prob. Median	15.35	15.98	14.72
Median Rainfall (mm)	Outcome	20.90	13.60	19.60
	Most Likely	20.00	20.00	18.00
	Prob. Median	13.75	15.00	12.50
Standard Deviation (mm)	Outcome	11.58	15.15	10.61
	Most Likely	6.99	5.38	6.92
	Prob. Median	6.25	5.16	7.84

Table 5.5.3.1.5 Correlation of Heavy Rainfall Warning forecasts with outcomes in Thames Region. (Rainfall Totals)

Correlation Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Correlation (dimensionless)	Most Likely	0.50	0.55	0.43
	Prob. Median	0.38	0.10	0.43
	$Const_{2mm/hr}$	0.31	0.41	0.38
Regression Coefficient (dimensionless)	Most Likely	0.82	1.54	0.66
	Prob. Median	0.71	0.30	0.58
	$Const_{2mm/hr}$	0.42	0.78	0.36
Correlation of Log-Rainfall (dimensionless)	Most Likely	0.41	0.43	0.42
	Prob. Median	0.30	0.28	0.34
	$Const_{2mm/hr}$	0.29	0.26	0.29
Regression Coeff. of Log-Rainfall (dimensionless)	Most Likely	0.84	2.24	0.65
	Prob. Median	0.52	1.18	0.36
	$Const_{2mm/hr}$	0.46	0.91	0.32

Table 5.5.3.1.6 Comparison of forecast sources: Standardised differences for assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Totals) (In this Table, the base forecast is "Most Likely")

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error	Prob. Median	0.96	-0.04	0.39
	$Const_{2mm/hr}$	1.13	0.33	0.57
	$Const_{20mm}$	0.58	0.69	0.21
Root Mean Square Error	Prob. Median	0.95	0.82	1.42
	$Const_{2mm/hr}$	0.83	0.34	1.30
	$Const_{20mm}$	0.73	1.00	0.47
Mean Absolute Error of Log-Rainfall	Prob. Median	1.16	-0.09	0.95
	$Const_{2mm/hr}$	1.24	0.27	0.61
	$Const_{20mm}$	0.48	0.71	0.02
Root Mean Square Error of Log-Rainfall	Prob. Median	0.45	-0.51	1.62
	$Const_{2mm/hr}$	0.61	-0.32	1.32
	$Const_{20mm}$	0.73	1.03	0.40

5.5.3.2 Assessment of category-forecasts

In addition to dealing with forecasts of rainfall amounts, Section 2.2.4 has outlined a number of measures of forecast performance appropriate for use where forecasts are in the form of simple statements as to whether or not a certain threshold will be exceeded. The forecasts provided by the Heavy Rainfall Warnings can be converted to be of this form and, since a number of different thresholds of rainfall amounts can be selected, they potentially provide a useful means of assessing the underlying forecasts' ability to distinguish between zero- and non-zero rainfall conditions and moderate and high-rainfall conditions. However, it should be noted that the present analysis applies only to the Warnings which were actually issued, and it does not present a complete picture of the performance of the Heavy Rainfall Warning Service.

Tables 5.5.3.2.1 to 5.5.3.2.2 show results for a collection of performance measures for analyses using thresholds of 20 and 25mm for the maximum rainfall accumulations in each of the 3 sub-areas of Thames Region. Results are given for the 4 types of forecasts listed in Table 5.5.1.1 and, in addition, results are given for what the values of the performance measures would be if forecasts of exceedences and non-exceedences of the threshold were made at random with the same rate of occurrence as found for the outcomes across all of the test occasions included in this study. The results for this type of forecast are listed against the name "Climatology": they provide a point of comparison for the candidate forecasts since a good forecast should do much better than the type of random forecast represented by "Climatology". For completeness, results are given for a second type of random forecast: these appear in parentheses after the actual values for the performance measure. In these cases, the random forecasts have a rate of forecasting threshold-exceedence equal to that observed for the actual forecasts.

The types of performance measures available for categorical forecasts fall naturally into two groups, and each table is divided in two corresponding parts. In the first group are the ordinary score statistics in which the performance measures are defined fairly directly in terms of the rates of occurrences of success or failure of the forecasts: these are listed in part **a** of each Table. The second group includes more refined measures in which the forecast performance is measured relative to what could be achieved by random forecasts of the two types outlined above: these are listed in part **b** of each Table.

In constructing the tables of results for performance measures of categorical forecasts, there are many cases where the values cannot be calculated because of the need to divide by zero: in these cases the results are represented by an asterisk (*). This rule has been applied even in cases where the standard formula formally gives 0/0 and where there is a potential to create a meaningful numerical value by re-expressing the formula in an alternative way.

In considering the results presented in the Tables 5.5.3.2.1-2, it should be recalled that these are based on relatively few forecast occasions compared to results reported for Evening Updates in Section 5.4.4. The values of the performance measures differ quite a lot between the sub-areas and this is a reflection of the small sample size. If the Relative Scores are used as the main basis of comparison, it can be seen that the simple forecast "$Const_{2mm/hr}$", where the forecast values are constructed to be

proportional to the interval length, is judged to perform best out of the candidate forecasts for the Western sub-area (where the number of forecast occasions for analysis is only 13). The sizes of the Relative Score statistics are moderately large in some instances, but this seems to be rather misleading since no account is taken of the very limited sample size.

Table 5.5.3.2.1a **Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. Rainfall Total > 20.0mm**

Ordinary Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
H Hit Rate	Most Likely	0.72 (0.50)	0.67 (0.52)	0.62 (0.51)
	Prob. Median	0.67 (0.50)	0.72 (0.61)	0.69 (0.53)
	$Const_{2mm/hr}$	0.50 (0.50)	0.72 (0.57)	0.77 (0.51)
	$Const_{20mm}$	0.50 (0.50)	0.67 (0.67)	0.54 (0.54)
	Climatology	0.50	0.56	0.50
CSI Critical Success Index	Most Likely	0.55 (0.31)	0.40 (0.24)	0.38 (0.27)
	Prob. Median	0.33 (0.14)	0.29 (0.13)	0.33 (0.13)
	$Const_{2mm/hr}$	0.25 (0.25)	0.38 (0.18)	0.57 (0.27)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.33	0.20	0.30
FAR False Alarm Rate	Most Likely	0.25 (0.50)	0.50 (0.67)	0.40 (0.54)
	Prob. Median	0.00 (0.50)	0.33 (0.67)	0.00 (0.54)
	$Const_{2mm/hr}$	0.50 (0.50)	0.40 (0.67)	0.20 (0.54)
	$Const_{20mm}$	* (*)	* (*)	* (*)
	Climatology	0.50	0.67	0.54
POD Probability of Detection	Most Likely	0.67 (0.44)	0.67 (0.44)	0.50 (0.38)
	Prob. Median	0.33 (0.17)	0.33 (0.17)	0.33 (0.15)
	$Const_{2mm/hr}$	0.33 (0.33)	0.50 (0.28)	0.67 (0.38)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.50	0.33	0.46
B Bias Ratio	Most Likely	0.89 (0.89)	1.33 (1.33)	0.83 (0.83)
	Prob. Median	0.33 (0.33)	0.50 (0.50)	0.33 (0.33)
	$Const_{2mm/hr}$	0.67 (0.67)	0.83 (0.83)	0.83 (0.83)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.5.3.2.1b **Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. Rainfall Total > 20.0mm**

Relative Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
HSS Heidke Skill Score	Most Likely	0.44 (0.00)	0.31 (0.00)	0.22 (0.00)
	Prob. Median	0.33 (0.00)	0.29 (0.00)	0.35 (0.00)
	$Const_{2mm/hr}$	0.00 (0.00)	0.35 (0.00)	0.53 (0.00)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.44 (0.00)	0.33 (0.00)	0.21 (0.00)
	Prob. Median	0.33 (0.00)	0.25 (0.00)	0.33 (0.00)
	$Const_{2mm/hr}$	0.00 (0.00)	0.33 (0.00)	0.52 (0.00)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.29 (0.00)	0.18 (0.00)	0.12 (0.00)
	Prob. Median	0.20 (0.00)	0.17 (0.00)	0.21 (0.00)
	$Const_{2mm/hr}$	0.00 (0.00)	0.21 (0.00)	0.36 (0.00)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	3.00 (1.00)	2.00 (1.00)	1.75 (1.00)
	Prob. Median	* (1.00)	4.00 (1.00)	* (1.00)
	$Const_{2mm/hr}$	1.00 (1.00)	3.00 (1.00)	4.67 (1.00)
	$Const_{20mm}$	* (1.00)	* (1.00)	* (1.00)
	Climatology	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	2.33 (1.00)	2.00 (1.00)	1.43 (1.00)
	Prob. Median	1.50 (1.00)	1.38 (1.00)	1.50 (1.00)
	$Const_{2mm/hr}$	1.00 (1.00)	1.67 (1.00)	2.57 (1.00)
	$Const_{20mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00
θ Odds Ratio	Most Likely	7.20 (1.00)	4.00 (1.00)	2.50 (1.00)
	Prob. Median	* (1.00)	5.50 (1.00)	* (1.00)
	$Const_{2mm/hr}$	1.00 (1.00)	5.00 (1.00)	12.00 (1.00)
	$Const_{20mm}$	* (1.00)	* (1.00)	* (1.00)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.5.3.2.2a **Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. Rainfall Total > 25.0mm**

Ordinary Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
H Hit Rate	Most Likely	0.67 (0.65)	0.83 (0.69)	0.54 (0.53)
	Prob. Median	0.78 (0.70)	0.78 (0.78)	0.62 (0.53)
	$Const_{2mm/hr}$	0.67 (0.67)	0.72 (0.69)	0.69 (0.51)
	$Const_{20mm}$	0.72 (0.72)	0.78 (0.78)	0.54 (0.54)
	Climatology	0.60	0.65	0.50
CSI Critical Success Index	Most Likely	0.14 (0.12)	0.40 (0.11)	0.14 (0.13)
	Prob. Median	0.20 (0.05)	0.00 (0.13)	0.17 (0.07)
	$Const_{2mm/hr}$	0.14 (0.14)	0.17 (0.11)	0.43 (0.23)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.16	0.12	0.30
FAR False Alarm Rate	Most Likely	0.67 (0.72)	0.33 (0.78)	0.50 (0.54)
	Prob. Median	0.00 (0.72)	* (*)	0.00 (0.54)
	$Const_{2mm/hr}$	0.67 (0.72)	0.67 (0.78)	0.25 (0.54)
	$Const_{20mm}$	* (*)	* (*)	* (*)
	Climatology	0.72	0.78	0.54
POD Probability of Detection	Most Likely	0.20 (0.17)	0.50 (0.17)	0.17 (0.15)
	Prob. Median	0.20 (0.06)	0.00 (0.00)	0.17 (0.08)
	$Const_{2mm/hr}$	0.20 (0.17)	0.25 (0.17)	0.50 (0.31)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.28	0.22	0.46
B Bias Ratio	Most Likely	0.60 (0.60)	0.75 (0.75)	0.33 (0.33)
	Prob. Median	0.20 (0.20)	0.00 (0.00)	0.17 (0.17)
	$Const_{2mm/hr}$	0.60 (0.60)	0.75 (0.75)	0.67 (0.67)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.5.3.2.2b **Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. Rainfall Total > 25.0mm**

Relative Scores

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
HSS Heidke Skill Score	Most Likely	0.05 (0.00)	0.47 (0.00)	0.03 (0.00)
	Prob. Median	0.27 (0.00)	0.00 (0.00)	0.18 (0.00)
	$Const_{2mm/hr}$	0.05 (0.00)	0.12 (0.00)	0.37 (0.00)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.05 (0.00)	0.43 (0.00)	0.02 (0.00)
	Prob. Median	0.20 (0.00)	0.00 (0.00)	0.17 (0.00)
	$Const_{2mm/hr}$	0.05 (0.00)	0.11 (0.00)	0.36 (0.00)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.03 (0.00)	0.31 (0.00)	0.01 (0.00)
	Prob. Median	0.15 (0.00)	0.00 (0.00)	0.10 (0.00)
	$Const_{2mm/hr}$	0.03 (0.00)	0.06 (0.00)	0.22 (0.00)
	$Const_{20mm}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	1.30 (1.00)	7.00 (1.00)	1.17 (1.00)
	Prob. Median	* (1.00)	* (*)	* (1.00)
	$Const_{2mm/hr}$	1.30 (1.00)	1.75 (1.00)	3.50 (1.00)
	$Const_{20mm}$	* (1.00)	* (1.00)	* (1.00)
	Climatology	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	1.06 (1.00)	1.86 (1.00)	1.03 (1.00)
	Prob. Median	1.25 (1.00)	1.00 (1.00)	1.20 (1.00)
	$Const_{2mm/hr}$	1.06 (1.00)	1.14 (1.00)	1.71 (1.00)
	$Const_{20mm}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00
θ Odds Ratio	Most Likely	1.37 (1.00)	13.00 (1.00)	1.20 (1.00)
	Prob. Median	* (1.00)	* (*)	* (1.00)
	$Const_{2mm/hr}$	1.37 (1.00)	2.00 (1.00)	6.00 (1.00)
	$Const_{20mm}$	* (1.00)	* (1.00)	* (1.00)
	Climatology	1.00	1.00	1.00

*	indicates scores which cannot be calculated because of zero-divisors.
( )	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.
"Climatology"	indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

5.5.4 Assessment of Probability Forecasts of Accumulations

One of the potentially important parts of the Heavy Rainfall Warning forecast is its probability forecast content. This section outlines an analysis which assesses how well the probability forecasts have performed. As for the assessment of category-forecasts, the analysis here takes account only of the Heavy Rainfall Warnings that were actually issued during the case-study periods.

The analysis here uses a performance measure appropriate to probability forecasts and compares the results found for the Heavy Rainfall Warning forecasts with certain other forecasts. The selected performance measure can equally-well be applied to single-valued forecasts, where the forecast is treated as expressing absolute certainty in a single value. In this case the performance measure is directly equivalent to the usual Mean Absolute Error statistic. Table 5.5.1.1 lists the single-valued forecasts used here. This is essentially the same set of forecasts used for the direct analysis of single-valued forecasts in Section 5.5.3. In addition, as outlined in Section 5.5.1 and Table 5.5.1.1, a set of probability forecasts have been created for comparison with those in the Heavy Rainfall Warning by taking the single-valued forecasts and attaching a somewhat arbitrary uncertainty-band: when the forecast amount is moderately large, this band extends from 0 up to twice the central forecast amount. The specification of this uncertainty band has not been subjected to detailed consideration and is simply put forward for comparison against the performance of the Heavy Rainfall Warning probability forecasts.

The results of the analysis of the probability forecasts are given in Table 5.5.4.1. The upper part of the table relates to the performance of the single-valued forecasts when treated as expressing absolute certainty. Values here are identical to those for the Mean Absolute Error given in Table 5.5.3.1.1 and they are repeated here because the Continuous Brier Score is identical to the Mean Absolute Error when a single-valued forecast is treated as absolutely certain. The lower part of the Table gives the Continuous Brier Score for the constructed probability forecasts and for the Heavy Rainfall Warnings' probability forecasts. It can be seen that including the uncertainty band with the single-valued forecasts has always decreased the performance measure in these cases. However, note that adding uncertainty of greater amounts would eventually lead to an increase in the score. The results for the Heavy Rainfall Warnings' probability forecasts are somewhat disappointing in comparison with those for the constructed probability forecasts, particularly when considering the probability forecast obtained from the "Most Likely" forecast by adding a simple uncertainty band. It seems that the probability forecasts contained in the Heavy Rainfall Warnings are no better than could be obtained by a simple uncertainty band centred about the main forecast-value. The same result was found for the Evening Update forecasts. It is interesting to note the performance of the probability forecast constructed by adding a 100% error to the "$Const_{20mm}$" forecast (corresponding to a fixed uniform distribution over 0-40mm): the results show that this is competitive with the operational forecasts in representing the uncertainty in the amount of rainfall that will fall once a Warning event has been identified and a forecast time-period has been determined.

Tables 5.5.4.2 and 5.5.4.3 relate directly to the question of whether there is enough evidence in the test dataset to distinguish between the performances of the different types of probability forecast. Taking the 'most likely' forecast ("Most Likely") from

the Heavy Rainfall Warning, with the addition of either zero or 100% uncertainty, as a "base forecast", Table 5.5.4.2 considers each of the other forecast sources in turn and asks how much evidence there is that the "base forecast" is better than the candidate forecast. In Table 5.5.4.3 the "base forecast" is the probability forecast contained in the Heavy Rainfall Warnings. The values in these tables are the standardised differences discussed earlier in Section 2.2.5, and positive values indicate that the "base forecast" has a better performance, as measured by the Continuous Brier Score, than the candidate. If the candidate forecast had a better performance, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the Continuous Brier Scores for the two forecast sources will turn out to be in the order indicated. For the purposes here, taking into account the small size of the data sample, a standardised difference outside the range ± 2.1 units indicates fairly strong evidence that one forecast source really is better than another.

The results shown in Tables 5.5.4.2 and 5.5.4.3 indicate that, if there are any differences in the performance of the probability forecasts, it is not strong enough to be revealed by this dataset.

Once again, any conclusion about the probability forecasts in the Heavy Rainfall Warnings needs to be tempered by the consideration that the probability forecasts in the Heavy Rainfall Warnings are not given to a high resolution and it may be that if a finer resolution had been used, better results might have been obtained.

Table 5.5.4.1 Assessment measures for Probability Forecasts associated with Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Totals)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error (mm)	*(certain)*			
	Most Likely	8.11	10.06	8.37
	Prob. Median	9.38	10.00	8.97
	$Const_{2mm/hr}$	10.29	10.63	9.25
	$Const_{20mm}$	8.99	10.93	8.75
Continuous Brier Score (mm)	*(100% error)*			
	Most Likely	6.19	7.13	6.28
	Prob. Median	7.02	7.90	8.00
	$Const_{2mm/hr}$	7.11	7.35	7.79
	$Const_{20mm}$	6.63	8.11	6.32
	(operational)			
	Prob. Forecast	6.62	7.50	7.02

Table 5.5.4.2 Comparison of forecast sources: Standardised Differences of Assessment measures for Probability Forecasts associated with Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Totals) (In this Table, the base forecast is "Most Likely" with either zero or 100% error)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Absolute Error (mm)	*(certain)*			
	Prob. Median	0.96	-0.04	0.39
	$Const_{2mm/hr}$	1.13	0.33	0.57
	$Const_{20mm}$	0.58	0.69	0.21
Continuous Brier Score (mm)	*(100% error)*			
	Prob. Median	0.90	0.51	1.23
	$Const_{2mm/hr}$	1.00	0.20	1.81
	$Const_{20mm}$	0.56	0.97	0.06
	(operational)			
	Prob. Forecast	0.59	0.27	0.78

Table 5.5.4.3 Comparison of forecast sources: Standardised Differences of Assessment measures for Probability Forecasts associated with Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Totals) (In this Table, the base forecast is "Prob. Forecast": the operational probability forecast)

Assessment Measure	Forecast Source	Area of Thames Region		
		NE	SE	W
Continuous Brier Score (mm)	*(100% error)*			
	Most Likely	-0.59	-0.27	-0.78
	Prob. Median	1.56	1.51	1.61
	$Const_{2mm/hr}$	0.44	-0.08	0.67
	$Const_{20mm}$	0.00	0.66	-0.55

5.5.5 Assessment of Single-valued Forecasts of Rates

5.5.5.1 Assessment of forecast rates

The analysis here for forecasts of rainfall rates in Heavy Rainfall Warnings follows the same outline as used in Section 5.5.3.1 for forecasts of rainfall amounts from the same source. An extra complication for the present case is that there are two potential sources of "ground-truth", deriving either from a network of raingauges or from weather radar. The analysis is also similar to that given for the forecasts of rainfall rates contained in the Evening Update forecasts. It should be recalled that the quantities being forecasted here relate to the maximum rainfall rate experienced at any time in the time-period of each Heavy Rainfall warning and at any location within a given sub-area of Thames Region of the Environment Agency.

Table 5.5.5.1.1 shows the basic assessment measures for the size of forecast errors for rainfall rates, evaluated for the 3 sub-areas of the Thames region. Results are given for the 4 types of forecasts listed in Table 5.5.1.2. As in earlier analyses of the same type, results are given for the best performance measure obtainable by a constant-value forecast (rows labelled "$Const_{best}$"). Table 5.5.5.1.2 shows the corresponding R^2 (efficiency) measures: these effectively compare the values of the performance measures shown in Table 5.5.5.1.1 with the best performance measure achievable by a constant-value forecast: that is, they compare the performance measures, as given in Table 5.5.5.1.1, for the given forecast source with the corresponding results for "$Const_{best}$" .

The results in Table 5.5.5.1.1 indicate that the forecasts contained within the Heavy Rainfall Warnings are considerably better matched to the ground-truth obtained from the raingauge network than they are to that from the weather radar source used here. The same result was found in Section 5.4.5.1 for the Evening Update forecasts of rainfall rates. Section 5.5.1 has outlined the potential problems with data from this radar source and results here are to be treated with caution. Examination of the example data in Tables 5.5.2.2 and 5.5.2.3 shows that the spatial maxima obtained from the radar source are typically much larger than those found from the raingauge network. Some of the difference in forecast performance between these sources that is shown in Table 5.5.5.1.1 arises from this fact.

The results for the raingauge-derived ground-truth in Tables 5.5.5.1.1 and 5.5.5.1.2 show that the operational forecasts for rainfall rates have rather similar performances to those of the simpler constant-valued forecasts. This is highlighted by Table 5.5.5.1.2, where the R^2 (efficiency) measures for the operational forecasts are usually negative. As was the case with forecasts of rainfall amounts (Section 5.5.3.1), the performance of the median value extracted from the probability forecasts is quite a lot worse than the "Most Likely" forecast, at least for the Northeast and Western sub-areas.

Table 5.5.5.1.3 shows details of the bias contained in the various forecast sources and compared with the two-versions of ground-truth. This clearly reveals the difference in the biases associated with the ground-truths. Overall it seems that the Heavy Rainfall Warning forecasts give values which are rather too small (compared with the

raingauge-derived ground truth), with the forecasts derived as the median of the probability forecasts tending to be somewhat smaller than the 'most likely' values. Table 5.5.5.1.4 shows some statistics for the rainfall amounts which give more details of the typical amounts obtained for the actual outcomes and for the forecasts. The results here show that the outcome values derived from radar have much higher standard-deviations than those derived from raingauges, as well as much larger means and medians. The larger means and standard deviations of the spatial maxima derived from radar compared with those from raingauges appears to arise from the finer spatial resolution of the former source. The operational forecasts are not well-tuned to either way of deriving the spatial maximum.

Table 5.5.5.1.5 gives values for correlation and regression coefficients for linear relationships between outcomes and forecasts of rainfall and log-rainfall. The interpretation of the coefficients here is similar to that outlined in the discussion of Evening Update forecasts in Section 5.4.5.1. The present case-study has included the "$\text{Const}_{4\text{mm/hr/hr}}$" forecast in which the forecast of the maximum rainfall rate is proportional to the interval-length. Table 5.5.5.1.4 shows that this type of forecast can have a positive correlation with the eventual outcome and this indicates that at least part of the skill of the operational forecasts of rainfall amounts arises from getting the event-length correct. Overall the correlations are not very large given the sample sizes available. There is some indication that the operational forecasts have a higher correlation with the radar-derived ground-truth than with that derived from raingauges. The corresponding result in the case of rainfall rates in the Evening Update forecasts was that the correlations were roughly the same size for the two different ground-truths.

Table 5.5.5.1.6 relates to the question of whether the evidence provided by the test dataset is sufficient to distinguish between the performance of different forecast sources, bearing in mind the sampling variability of the forecast performance statistics and the statistical dependences between them. Taking the 'most likely' forecast ("Most Likely") from the Heavy Rainfall Warnings as a "base forecast", Table 5.5.5.1.6 considers each of the other forecast sources in turn and asks how much evidence there is that the "base forecast" is better than the candidate forecast. Once again, the values given are the standardised differences discussed earlier in Section 4.3.3, and positive values indicate that the "base forecast" has a smaller size of error, as measured by the performance statistic, than the candidate. If the candidate forecast produces smaller errors, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the long-run performance measures for the two forecast sources will turn out to be in the order indicated. For the purposes here a standardised difference outside the range ± 2 units (or ± 2.1 units to allow for the small sample size) indicates fairly strong evidence that one forecast source really is better than another.

The results in Table 5.5.5.1.6 reflect those in Table 5.5.5.1.1, in that the comparisons which favour one forecast source over another are the same. However, Table 5.5.5.1.6 provides extra information. For example, it shows that only for the Southeast sub-area is there strong weak evidence that the 'most likely' values in the Heavy Rainfall Warning provide better forecasts of the raingauge-derived ground-truth than the median-values derived from the probability forecasts. However, this preference is consistent across the sub-areas and performance measures, and thus the analysis does

raise a doubt over the construction of the probability forecast, or at least over its resolution in terms of rainfall-rate intervals. The disparity in performance of the median of the probability forecast and the 'Most Likely' forecast appears greater here than in other analyses, and thus there may be some special problem relating to the interpretation of the probability tables.

Table 5.5.5.1.1 Raw assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Rates)

Assessment Measure	Forecast Source	Raingauge Ground-Truth Area of Thames Region			Radar Ground-Truth Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error (mm/hr)	Most Likely	18.54	14.76	16.05	60.71	49.29	59.43
	Prob. Median	20.60	17.88	16.29	60.99	51.01	60.62
	$Const_{4mm/hr/hr}$	20.78	20.87	25.60	47.44	41.52	49.74
	$Const_{30mm/hr}$	16.00	16.80	14.25	52.00	52.52	51.67
	$Const_{best}$	16.00	14.98	12.68	38.43	35.79	44.00
Root Mean Square Error (mm/hr)	Most Likely	23.35	20.63	19.59	76.48	63.60	76.91
	Prob. Median	26.12	24.01	22.29	77.07	66.11	77.78
	$Const_{4mm/hr/hr}$	24.65	26.25	32.61	62.14	52.16	63.98
	$Const_{30mm/hr}$	19.83	20.21	16.58	69.33	58.80	68.80
	$Const_{best}$	19.78	19.59	16.00	49.79	45.66	52.27
Mean Absolute Error of Log-Rainfall (dim'less)	Most Likely	0.77	0.78	0.69	1.36	1.15	1.30
	Prob. Median	0.90	0.94	0.69	1.40	1.26	1.36
	$Const_{4mm/hr/hr}$	0.79	0.98	0.81	0.82	0.82	0.90
	$Const_{30mm/hr}$	0.66	0.88	0.58	0.89	0.80	0.93
	$Const_{best}$	0.66	0.81	0.52	0.58	0.62	0.74
Root Mean Square Error of Log-Rainfall (dim'less)	Most Likely	0.91	1.19	0.76	1.49	1.25	1.48
	Prob. Median	1.08	1.34	0.86	1.56	1.40	1.55
	$Const_{4mm/hr/hr}$	1.04	1.41	1.00	0.95	0.92	1.04
	$Const_{30mm/hr}$	0.89	1.30	0.69	1.03	0.95	1.06
	$Const_{best}$	0.85	1.15	0.60	0.74	0.78	0.87

Table 5.5.5.1.2 R^2 (efficiency) measures for Heavy Rainfall Warning forecasts in the Thames Region for each type of assessment measure. (Rainfall Rates)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
R^2 for Mean Absolute Error	Most Likely	-0.16	0.01	-0.27	-0.58	-0.38	-0.35
	Prob. Median	-0.29	-0.19	-0.28	-0.59	-0.43	-0.38
	$Const_{4mm/hr/hr}$	-0.30	-0.39	-1.02	-0.23	-0.16	-0.13
	$Const_{30mm/hr}$	0.00	-0.12	-0.12	-0.35	-0.19	-0.17
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00
R^2 for Root Mean Square Error	Most Likely	-0.39	-0.11	-0.50	-1.36	-0.94	-1.16
	Prob. Median	-0.74	-0.50	-0.94	-1.40	-1.10	-1.21
	$Const_{4mm/hr/hr}$	-0.55	-0.80	-3.15	-0.56	-0.30	-0.50
	$Const_{30mm/hr}$	-0.01	-0.06	-0.07	-0.94	-0.66	-0.73
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00
R^2 for Mean Absolute Error of Log-Rainfall	Most Likely	-0.18	0.03	-0.33	-1.34	-0.87	-0.77
	Prob. Median	-0.37	-0.16	-0.33	-1.42	-1.04	-0.85
	$Const_{4mm/hr/hr}$	-0.21	-0.21	-0.56	-0.41	-0.33	-0.22
	$Const_{30mm/hr}$	0.00	-0.09	-0.12	-0.53	-0.29	-0.27
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00
R^2 for Root Mean Square Error of Log-Rainfall	Most Likely	-0.13	-0.05	-0.61	-3.07	-1.58	-1.88
	Prob. Median	-0.59	-0.34	-1.10	-3.49	-2.25	-2.14
	$Const_{4mm/hr/hr}$	-0.49	-0.50	-1.83	-0.66	-0.40	-0.41
	$Const_{30mm/hr}$	-0.08	-0.28	-0.32	-0.96	-0.47	-0.46
	$Const_{best}$	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.5.5.1.3 Bias measures for Heavy Rainfall Warning forecasts in the Thames Region . (Rainfall Rates)

Bias Measure	Forecast Source	Raingauge Ground-Truth Area of Thames Region			Radar Ground-Truth Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Error (mm/hr)	Most Likely	12.52	5.02	6.20	59.30	47.05	55.27
	Prob. Median	12.46	6.41	6.96	59.24	48.44	56.02
	$Const_{4mm/hr/hr}$	-6.87	-12.87	-20.18	39.91	29.16	28.88
	$Const_{30mm/hr}$	1.47	-4.98	-4.34	48.25	37.05	44.73
Median Error (mm/hr)	Most Likely	12.60	1.70	8.00	54.52	32.86	55.53
	Prob. Median	11.60	2.92	4.00	49.50	34.28	51.53
	$Const_{4mm/hr/hr}$	-8.80	-14.00	-21.60	32.33	26.22	18.62
	$Const_{30mm/hr}$	-1.20	-12.00	-11.60	42.33	18.63	33.53
Mean Error of Log-Rainfall (dim'less)	Most Likely	0.35	-0.10	0.24	1.30	1.04	1.18
	Prob. Median	0.38	0.00	0.28	1.34	1.14	1.21
	$Const_{4mm/hr/hr}$	-0.39	-0.76	-0.66	0.57	0.38	0.27
	$Const_{30mm/hr}$	-0.24	-0.61	-0.34	0.72	0.53	0.59
Median Error of Log-Rainfall (dim'less)	Most Likely	0.34	0.11	0.52	1.21	1.09	1.52
	Prob. Median	0.37	0.20	0.26	1.28	0.93	1.13
	$Const_{4mm/hr/hr}$	-0.16	-0.47	-0.62	0.63	0.63	0.50
	$Const_{30mm/hr}$	-0.04	-0.51	-0.49	0.88	0.49	0.75

Table 5.5.5.1.4 Statistics of forecasts and outcomes for Heavy Rainfall Warning forecasts in Thames Region. (Rainfall Rates)

Statistic of Rainfall	Forecast Source	Area of Thames Region		
		NE	SE	W
Mean Rainfall (mm/hr)	Outcome (Raingauge)	31.47	25.02	25.66
	Outcome (Radar)	78.25	67.05	74.73
	Most Likely	18.94	20.00	19.46
	Prob. Median	19.00	18.61	18.71
Median Rainfall (mm/hr)	Outcome (Raingauge)	28.80	18.00	18.40
	Outcome (Radar)	72.33	48.83	63.53
	Most Likely	17.50	20.00	20.00
	Prob. Median	17.00	17.50	15.00
Standard Deviation (mm/hr)	Outcome (Raingauge)	20.35	20.16	16.66
	Outcome (Radar)	51.24	46.98	54.41
	Most Likely	8.94	8.35	11.24
	Prob. Median	10.71	9.68	11.55

Table 5.5.5.1.5 Correlation of Heavy Rainfall Warning forecasts with outcomes in Thames Region. (Rainfall Rates)

Correlation Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Correlation	Most Likely	0.23	0.16	0.08	0.26	0.43	-0.01
	Prob. Median	-0.07	-0.17	-0.20	0.15	0.17	-0.05
	$Const_{4mm/hr/hr}$	0.15	0.16	0.07	0.29	0.32	-0.04
Regression Coefficient	Most Likely	0.52	0.38	0.12	1.46	2.43	-0.05
	Prob. Median	-0.13	-0.35	-0.28	0.73	0.84	-0.22
	$Const_{4mm/hr/hr}$	0.19	0.20	0.05	0.89	0.96	-0.09
Correlation of Log-Rainfall	Most Likely	0.33	0.15	0.23	0.38	0.48	0.27
	Prob. Median	0.05	-0.17	-0.04	0.26	0.23	0.13
	$Const_{4mm/hr/hr}$	-0.01	0.07	-0.02	0.24	0.08	-0.04
Regression Coef. of Log-Rainfall	Most Likely	0.55	0.37	0.25	0.55	0.77	0.42
	Prob. Median	0.07	-0.39	-0.05	0.33	0.36	0.22
	$Const_{4mm/hr/hr}$	-0.02	0.22	-0.03	0.41	0.15	-0.09

Table 5.5.5.1.6 Comparison of forecast sources: Standardised differences for assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Rates) (In this Table, the base forecast is "Most Likely")

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error	Prob. Median	1.53	2.92	0.13	0.19	1.36	0.65
	$Const_{4mm/hr/hr}$	0.53	1.14	1.61	-2.76	-1.68	-1.30
	$Const_{30mm/hr}$	-0.82	0.77	-0.59	-4.09	-3.65	-2.41
Root Mean Square Error	Prob. Median	1.43	2.55	1.19	0.41	1.48	0.44
	$Const_{4mm/hr/hr}$	0.28	0.82	1.57	-2.42	-1.92	-1.97
	$Const_{30mm/hr}$	-1.39	-0.15	-1.07	-3.45	-3.11	-3.40
Mean Absolute Error of Log-Rainfall	Prob. Median	1.69	2.86	0.01	0.53	1.55	0.53
	$Const_{4mm/hr/hr}$	0.09	1.05	0.78	-3.74	-2.67	-1.77
	$Const_{30mm/hr}$	-0.66	0.69	-0.91	-4.31	-3.98	-2.43
Root Mean Square Error of Log-Rainfall	Prob. Median	1.92	2.17	0.91	0.80	1.71	0.52
	$Const_{4mm/hr/hr}$	0.67	1.63	1.51	-3.53	-2.31	-2.18
	$Const_{30mm/hr}$	-0.11	1.02	-0.60	-3.62	-3.41	-3.01

5.5.5.2 Assessment of category-forecasts

The analysis here for the Heavy Rainfall Warning forecasts of whether rainfall rates will exceed given thresholds follows the same outline as that used in Section 5.5.3.2 for forecasts of rainfall amount. There are two potential sources of "ground-truth", deriving either from a network of raingauges or from weather radar. As in Section 5.5.3.2, the tables of results include values for the performance measures that would be achieved by two types of random forecast, one based on the observed rate of threshold-exceedence among the outcomes for the given ground-truth and one based on the rate found for the given forecast source.

Tables 5.5.5.2.1 and 5.5.5.2.2 show results for a collection of performance measures for analyses using thresholds of 15 and 25mm h^{-1} for the maximum rainfall rate in the forecast period of each Warning in each of the 3 sub-areas of Thames Region. The types of performance measures available for categorical forecasts fall naturally into two groups, and each table is divided into two corresponding parts. In the first group are the "Ordinary Score" statistics in which the performance measures are defined fairly directly in terms of the rates of occurrences of success or failure of the forecasts: these are listed in part **a** of each Table. The second group, termed "Relative Scores", includes more refined measures in which the forecast performance is measured relative to what could be achieved by random forecasts of the two types outlined above: these are listed in part **b** of each Table.

In constructing the tables of results for performance measures of categorical forecasts, there are many cases where the values cannot be calculated because of the need to divide by zero: in these cases the results are represented by an asterisk (*). This rule has been applied even in cases where the standard formula formally gives 0/0 and where there is a potential to create a meaningful numerical value by re-expressing the formula in an alternative way.

In considering the results presented in the Tables 5.5.5.2.1-2, it should be recalled that these are based on relatively few forecast occasions compared to results reported for Evening Updates in Section 5.4.5. The results are potentially badly affected by the small sample sizes, but they mainly suggest that there is no skill in the operational forecasts of rainfall rates. While some quite high relative skill scores are found for the Kuipers Skill Score comparing the operational forecasts against the radar-derived ground-truth at a threshold of 15 mm h^{-1}, this apparent skill disappears when the threshold is raised to 25 mm h^{-1}. Thus it is likely that this is simply an artifact arising from the small sample size.

Table 5.5.5.2.1a Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region. Rainfall Rate > 15.0mm h^{-1} Ordinary Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
H Hit Rate	Most Likely	0.50 (0.50)	0.50 (0.54)	0.46 (0.51)	0.56 (0.50)	0.61 (0.50)	0.69 (0.53)
	Prob. Median	0.50 (0.50)	0.39 (0.50)	0.38 (0.49)	0.56 (0.50)	0.50 (0.50)	0.62 (0.47)
	$Const_{4mm/hr/hr}$	0.72 (0.75)	0.67 (0.67)	0.62 (0.62)	0.89 (0.90)	0.89 (0.89)	0.85 (0.85)
	$Const_{30mm/hr}$	0.78 (0.78)	0.67 (0.67)	0.62 (0.62)	0.94 (0.94)	0.89 (0.89)	0.85 (0.85)
	Climatology	0.65	0.56	0.53	0.90	0.80	0.74
CSI Critical Success Index	Most Likely	0.44 (0.44)	0.44 (0.47)	0.36 (0.40)	0.53 (0.49)	0.59 (0.57)	0.64 (0.59)
	Prob. Median	0.44 (0.44)	0.31 (0.40)	0.27 (0.36)	0.53 (0.49)	0.47 (0.47)	0.55 (0.43)
	$Const_{4mm/hr/hr}$	0.72 (0.74)	0.67 (0.67)	0.62 (0.62)	0.89 (0.89)	0.89 (0.89)	0.85 (0.85)
	$Const_{30mm/hr}$	0.78 (0.78)	0.67 (0.67)	0.62 (0.62)	0.94 (0.94)	0.89 (0.89)	0.85 (0.85)
	Climatology	0.64	0.50	0.44	0.89	0.80	0.73
FAR False Alarm Rate	Most Likely	0.22 (0.22)	0.36 (0.33)	0.43 (0.38)	0.00 (0.06)	0.09 (0.11)	0.00 (0.15)
	Prob. Median	0.22 (0.22)	0.44 (0.33)	0.50 (0.38)	0.00 (0.06)	0.11 (0.11)	0.00 (0.15)
	$Const_{4mm/hr/hr}$	0.24 (0.22)	0.33 (0.33)	0.38 (0.38)	0.06 (0.06)	0.11 (0.11)	0.15 (0.15)
	$Const_{30mm/hr}$	0.22 (0.22)	0.33 (0.33)	0.38 (0.38)	0.06 (0.06)	0.11 (0.11)	0.15 (0.15)
	Climatology	0.22	0.33	0.38	0.06	0.11	0.15
POD Probability of Detection	Most Likely	0.50 (0.50)	0.58 (0.61)	0.50 (0.64)	0.53 (0.50)	0.62 (0.61)	0.64 (0.54)
	Prob. Median	0.50 (0.50)	0.42 (0.50)	0.38 (0.46)	0.53 (0.50)	0.50 (0.50)	0.55 (0.46)
	$Const_{4mm/hr/hr}$	0.93 (0.94)	1.00 (1.00)	1.00 (1.00)	0.94 (0.94)	1.00 (1.00)	1.00 (1.00)
	$Const_{30mm/hr}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	0.78	0.67	0.67	0.94	0.89	0.85
B Bias Ratio	Most Likely	0.64 (0.64)	0.92 (0.92)	0.88 (0.88)	0.53 (0.53)	0.69 (0.69)	0.64 (0.64)
	Prob. Median	0.64 (0.64)	0.75 (0.75)	0.75 (0.75)	0.53 (0.53)	0.56 (0.56)	0.55 (0.55)
	$Const_{4mm/hr/hr}$	1.21 (1.21)	1.50 (1.50)	1.62 (1.62)	1.00 (1.00)	1.12 (1.12)	1.18 (1.18)
	$Const_{30mm/hr}$	1.29 (1.29)	1.50 (1.50)	1.62 (1.62)	1.06 (1.06)	1.12 (1.12)	1.18 (1.18)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.5.5.2.1b Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region.
Rainfall Rate > 15.0mm h^{-1} Relative Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
HSS Heidke Skill Score	Most Likely	0.00 (0.00)	-0.09(0.00)	-0.10(0.00)	0.11 (0.00)	0.06 (0.00)	0.35 (0.00)
	Prob. Median	0.00 (0.00)	-0.22(0.00)	-0.21(0.00)	0.11 (0.00)	0.00 (0.00)	0.27 (0.00)
	$Const_{4mm/hr/hr}$	-0.10(0.00)	-0.00(0.00)	-0.00(0.00)	-0.06(0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{30mm/hr}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.00 (0.00)	-0.08(0.00)	-0.10(0.00)	0.53 (0.00)	0.12 (0.00)	0.64 (0.00)
	Prob. Median	0.00 (0.00)	-0.25(0.00)	-0.22(0.00)	0.53 (0.00)	0.00 (0.00)	0.55 (0.00)
	$Const_{4mm/hr/hr}$	-0.07(0.00)	-0.00(0.00)	-0.00(0.00)	-0.06(0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{30mm/hr}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.00 (0.00)	-0.04(0.00)	-0.05(0.00)	0.06 (0.00)	0.03 (0.00)	0.21 (0.00)
	Prob. Median	0.00 (0.00)	-0.10(0.00)	-0.09(0.00)	0.06 (0.00)	0.00 (0.00)	0.16 (0.00)
	$Const_{4mm/hr/hr}$	-0.05(0.00)	-0.00(0.00)	0.00 (0.00)	-0.03(0.00)	0.00 (0.00)	0.00 (0.00)
	$Const_{30mm/hr}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	1.00 (1.00)	0.88 (1.00)	0.83 (1.00)	* (*)	1.25 (1.00)	* (1.00)
	Prob. Median	1.00 (1.00)	0.62 (1.00)	0.62 (1.00)	* (*)	1.00 (1.00)	* (1.00)
	$Const_{4mm/hr/hr}$	0.93 (1.00)	1.00 (1.00)	1.00 (1.00)	0.94 (1.00)	1.00 (1.00)	1.00 (1.00)
	$Const_{30mm/hr}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	1.00 (1.00)	0.80 (1.00)	0.80 (1.00)	1.33 (1.00)	1.33 (1.00)	2.75 (1.00)
	Prob. Median	1.00 (1.00)	0.57 (1.00)	0.64 (1.00)	1.00 (1.00)	1.00 (1.00)	2.20 (1.00)
	$Const_{4mm/hr/hr}$	0.00 (1.00)	* (*)	* (*)	0.00 (1.00)	* (*)	* (*)
	$Const_{30mm/hr}$	* (*)	* (*)	* (*)	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
θ Odds Ratio	Most Likely	1.00 (1.00)	0.70 (1.00)	0.67 (1.00)	* (*)	1.67 (1.00)	* (1.00)
	Prob. Median	1.00 (1.00)	0.36 (1.00)	0.40 (1.00)	* (*)	1.00 (1.00)	* (1.00)
	$Const_{4mm/hr/hr}$	0.00 (1.00)	* (*)	* (*)	0.00 (1.00)	* (*)	* (*)
	$Const_{30mm/hr}$	* (*)	* (*)	* (*)	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00	*	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.5.5.2.2a Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region.
Rainfall Rate > 25.0mm h^{-1} Ordinary Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
H Hit Rate	Most Likely	0.44 (0.41)	0.56 (0.54)	0.38 (0.53)	0.22 (0.28)	0.28 (0.31)	0.23 (0.31)
	Prob. Median	0.44 (0.44)	0.44 (0.50)	0.46 (0.52)	0.33 (0.37)	0.28 (0.31)	0.31 (0.36)
	$Const_{4mm/hr/hr}$	0.44 (0.59)	0.39 (0.67)	0.31 (0.47)	0.67 (0.72)	0.56 (0.65)	0.62 (0.69)
	$Const_{30mm/hr}$	0.67 (0.67)	0.44 (0.67)	0.46 (0.46)	0.89 (0.89)	0.83 (0.83)	0.77 (0.77)
	Climatology	0.56	0.51	0.50	0.80	0.72	0.64
CSI Critical Success Index	Most Likely	0.23 (0.20)	0.20 (0.17)	0.00 (0.13)	0.18 (0.22)	0.19 (0.21)	0.09 (0.15)
	Prob. Median	0.29 (0.29)	0.09 (0.17)	0.12 (0.18)	0.29 (0.32)	0.19 (0.21)	0.18 (0.22)
	$Const_{4mm/hr/hr}$	0.44 (0.56)	0.31 (0.38)	0.31 (0.43)	0.67 (0.71)	0.56 (0.63)	0.62 (0.67)
	$Const_{30mm/hr}$	0.67 (0.67)	0.44 (0.44)	0.46 (0.46)	0.89 (0.89)	0.83 (0.83)	0.77 (0.77)
	Climatology	0.50	0.29	0.30	0.80	0.71	0.62
FAR False Alarm Rate	Most Likely	0.25 (0.33)	0.50 (0.56)	1.00 (0.54)	0.25 (0.11)	0.25 (0.17)	0.50 (0.23)
	Prob. Median	0.33 (0.33)	0.75 (0.56)	0.67 (0.54)	0.17 (0.11)	0.25 (0.17)	0.33 (0.23)
	$Const_{4mm/hr/hr}$	0.43 (0.33)	0.62 (0.56)	0.64 (0.54)	0.14 (0.11)	0.23 (0.17)	0.27 (0.23)
	$Const_{30mm/hr}$	0.33 (0.33)	0.56 (0.33)	0.54 (0.54)	0.11 (0.11)	0.17 (0.17)	0.23 (0.23)
	Climatology	0.33	0.33	0.54	0.11	0.17	0.23
POD Probability of Detection	Most Likely	0.25 (0.22)	0.25 (0.22)	0.00 (0.15)	0.19 (0.22)	0.20 (0.22)	0.10 (0.15)
	Prob. Median	0.33 (0.33)	0.12 (0.22)	0.17 (0.23)	0.31 (0.33)	0.20 (0.22)	0.20 (0.23)
	$Const_{4mm/hr/hr}$	0.67 (0.78)	0.62 (0.72)	0.67 (0.85)	0.75 (0.78)	0.67 (0.72)	0.80 (0.85)
	$Const_{30mm/hr}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	0.67	0.44	0.67	0.89	0.83	0.85
B Bias Ratio	Most Likely	0.33 (0.33)	0.50 (0.50)	0.33 (0.33)	0.25 (0.53)	0.27 (0.27)	0.20 (0.20)
	Prob. Median	0.50 (0.50)	0.50 (0.50)	0.50 (0.50)	0.38 (0.53)	0.27 (0.27)	0.30 (0.30)
	$Const_{4mm/hr/hr}$	1.17 (1.17)	1.62 (1.62)	1.83 (1.83)	0.88 (1.00)	0.87 (0.87)	1.10 (1.10)
	$Const_{30mm/hr}$	1.50 (1.50)	2.25 (2.25)	2.17 (2.17)	1.12 (1.06)	1.20 (1.20)	1.30 (1.30)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

Table 5.5.5.2.2b Categorical assessment measures for Heavy Rainfall Warning forecasts in the Thames Region.
Rainfall Rate > 25.0mm h^{-1} Relative Scores

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
HSS Heidke Skill Score	Most Likely	0.06 (0.00)	0.05 (0.00)	-0.30(0.00)	-0.09(0.00)	-0.05(0.00)	-0.12(0.00)
	Prob. Median	0.00 (0.00)	-0.18(0.00)	-0.12(0.00)	-0.06(0.00)	-0.05(0.00)	-0.07(0.00)
	$Const_{4mm/hr/hr}$	-0.36(0.00)	-0.16(0.00)	-0.31(0.00)	-0.17(0.00)	-0.26(0.00)	-0.23(0.00)
	$Const_{30mm/hr}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
KSS Kuipers Skill Score	Most Likely	0.08 (0.00)	0.05 (0.00)	-0.29(0.00)	-0.31(0.00)	-0.13(0.00)	-0.23(0.00)
	Prob. Median	0.00 (0.00)	-0.17(0.00)	-0.12(0.00)	-0.19(0.00)	-0.13(0.00)	-0.13(0.00)
	$Const_{4mm/hr/hr}$	-0.33(0.00)	-0.17(0.00)	-0.33(0.00)	-0.25(0.00)	-0.33(0.00)	-0.20(0.00)
	$Const_{30mm/hr}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
ETS Equitable Skill Score	Most Likely	0.03 (0.00)	0.03 (0.00)	-0.13(0.00)	-0.04(0.00)	-0.03(0.00)	-0.06(0.00)
	Prob. Median	0.00 (0.00)	-0.08(0.00)	-0.06(0.00)	-0.03(0.00)	-0.03(0.00)	-0.04(0.00)
	$Const_{4mm/hr/hr}$	-0.15(0.00)	-0.08(0.00)	-0.14(0.00)	-0.08(0.00)	-0.12(0.00)	-0.10(0.00)
	$Const_{30mm/hr}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Climatology	0.00	0.00	0.00	0.00	0.00	0.00
LR2 Likelihood Ratio for Above Threshold Occasions	Most Likely	1.50 (1.00)	1.25 (1.00)	0.00 (1.00)	0.38 (1.00)	0.60 (1.00)	0.30 (1.00)
	Prob. Median	1.00 (1.00)	0.42 (1.00)	0.58 (1.00)	0.62 (1.00)	0.60 (1.00)	0.60 (1.00)
	$Const_{4mm/hr/hr}$	0.67 (1.00)	0.78 (1.00)	0.67 (1.00)	0.75 (1.00)	0.67 (1.00)	0.80 (1.00)
	$Const_{30mm/hr}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
LR1 Likelihood Ratio for Below Threshold Occasions	Most Likely	1.11 (1.00)	1.07 (1.00)	0.71 (1.00)	0.62 (1.00)	0.83 (1.00)	0.74 (1.00)
	Prob. Median	1.00 (1.00)	0.80 (1.00)	0.86 (1.00)	0.73 (1.00)	0.83 (1.00)	0.83 (1.00)
	$Const_{4mm/hr/hr}$	0.00 (1.00)	0.53 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)
	$Const_{30mm/hr}$	* (*)	* (*)	* (*)	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00
θ Odds Ratio	Most Likely	1.67 (1.00)	1.33 (1.00)	0.00 (1.00)	0.23 (1.00)	0.50 (1.00)	0.22 (1.00)
	Prob. Median	1.00 (1.00)	0.33 (1.00)	0.50 (1.00)	0.45 (1.00)	0.50 (1.00)	0.50 (1.00)
	$Const_{4mm/hr/hr}$	0.00 (1.00)	0.42 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)
	$Const_{30mm/hr}$	* (*)	* (*)	* (*)	* (*)	* (*)	* (*)
	Climatology	1.00	1.00	1.00	1.00	1.00	1.00

* indicates scores which cannot be calculated because of zero-divisors.

( ) indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the forecast source.

"Climatology" indicates average score for a forecast generated randomly to forecast "above threshold" at the same rate as the observed outcomes.

5.5.6 Assessment of Probability Forecasts of Rates

This section outlines an analysis which assesses the performance of the probability forecasts for maximum rainfall rates which are part of the Heavy Rainfall Warning forecasts. The analysis is directly comparable to that employed for probability forecasts of maximum rainfall amounts reported in Section 5.5.4.

As for the other case where probability forecasts are being compared, two sets of forecasts are used for comparison against the probability forecasts contained in the Heavy Rainfall Warnings. As outlined in Section 5.5.1 and Table 5.5.1.2, one set of probability forecasts have been created from the set of single-valued forecasts by treating these as expressing complete certainty in the quoted value, and a second set has been created by taking each single-valued forecast and attaching a somewhat arbitrary uncertainty-band: when the forecast maximum rate is moderately large, this band extends from 0 up to twice the central forecast rate. The specification of this uncertainty band has not been subjected to detailed consideration and is simply put forward for comparison against the performance of the Heavy Rainfall Warning probability forecasts.

As for the other analyses of rainfall rates, two different versions of ground-truth are available, and results are given here for both.

The results of the analysis of the probability forecasts are given in Table 5.5.6.1. The upper part of the table relates to the performance of the single-valued forecasts when treated as expressing absolute certainty. Values here for the "certain" forecasts are identical to those for the Mean Absolute Error given in Table 5.5.5.1.1 and they are repeated here for comparison with the results of the other forecasts which do contain uncertainty. The lower part of the Table gives the Continuous Brier Score for the constructed probability forecasts and for the Heavy Rainfall Warnings' probability forecasts. It can be seen that including the uncertainty band with the single-valued forecasts has always decreased the performance measure in these cases. However, note that adding uncertainty of greater amounts would eventually lead to an increase in the score.

As for the analysis of the single-valued forecasts in Section 5.5.5.1, Table 5.5.6.1 again shows that the forecasts (and in particular the probability forecast) contained in the Heavy Rainfall Warnings are a better match to the outcomes derived from the raingauge network than they are to those obtained from the Nimrod radar-rainfall source.

In Table 5.5.6.1, the results for the Heavy Rainfall Warnings' probability forecasts are again somewhat disappointing in comparison with those for the constructed probability forecasts, particularly when considering the probability forecast obtained from the "Most Likely" forecast by adding a simple uncertainty band. It seems that the probability forecasts contained in the Heavy Rainfall Warnings are no better than could be obtained by a simple uncertainty band centred about the main forecast-value. The same result was found for the Evening Update forecasts. It is interesting to note the performance of the probability forecast constructed by adding a 100% error to the "$Const_{30mm/hr}$" forecast (which corresponds to a fixed uniform distribution over 0-60

mm h^{-1}): the results show that this is competitive with the operational forecasts in representing the uncertainty in the amount of rainfall that will fall once a Warning event has been identified and a forecast time-period has been determined.

Tables 5.5.6.2 and 5.5.6.3 relate directly to the question of whether there is enough evidence in the test dataset to distinguish between the performances of the different types of probability forecast. Taking the 'most likely' forecast ("Most Likely") from the Heavy Rainfall Warning, with the addition of either zero or 100% uncertainty, as a "base forecast", Table 5.5.6.2 considers each of the other forecast sources in turn and asks how much evidence there is that the "base forecast" is better than the candidate forecast. In Table 5.5.6.3 the "base forecast" is the probability forecast contained in the Heavy Rainfall Warnings. The values in these tables are the standardised differences discussed earlier in Section 4.3.3, and positive values indicate that the "base forecast" has a better performance, as measured by the Continuous Brier Score, than the candidate. If the candidate forecast had a better performance, then the value would be negative. The absolute size of the standardised difference indicates the strength of the evidence in the data that the Continuous Brier Scores for the two forecast sources will turn out to be in the order indicated. For the purposes here, taking account of the small size of the test dataset, a standardised difference outside the range ± 2.1 units indicates fairly strong evidence that one forecast source really is better than another.

The results shown in Tables 5.5.6.2 and 5.5.6.3 can be interpreted as follows for the case of the raingauge-based ground-truth There is no clear evidence to prefer any of the probability forecasts, including those derived from the constant-valued forecasts. An exception arises in the case of the 100% uncertainty band added to the median value of the operational probability forecast: this is close to having been shown to be worse than the operational probability forecast. In the case of the radar-based ground-truth, the extent of the mismatch in the values produced as forecasts in the Heavy Rainfall Warnings and those actually observed in the radar data is such that none of the probability forecasts are good. However there is clear evidence that the probability forecast consisting of a uniform distribution over the range 0-60 mm h^{-1} (i.e. "$Const_{30mm/hr}$" with 100% error band) is better than the operational forecasts when treated as a forecast of the maximum rainfall rate derived from the radar data.

As before, the conclusion about the probability forecasts for rainfall rates in the Heavy Rainfall Warnings needs to be tempered by the consideration that better results might have been obtained if the probability forecasts had been produced at a finer resolution.

Table 5.5.6.1 Assessment measures for Probability Forecasts associated with Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Rates)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error (mm/hr)	*(certain)*						
	Most Likely	18.54	14.76	16.05	60.71	49.29	59.43
	Prob. Median	20.60	17.88	16.29	60.99	51.01	60.62
	$Const_{4mm/hr/hr}$	20.78	20.87	25.60	47.44	41.52	49.74
	$Const_{30mm/hr}$	16.00	16.80	14.25	52.00	42.52	51.67
Continuous Brier Score (mm/hr)	*(100% error)*						
	Most Likely	14.41	11.03	10.38	53.66	42.26	53.72
	Prob. Median	16.23	13.66	12.38	54.06	43.90	74.25
	$Const_{4mm/hr/hr}$	14.11	14.24	16.20	36.45	30.39	37.90
	$Const_{30mm/hr}$	11.34	11.58	9.81	43.76	35.14	43.21
	(operational)						
	Prob. Forecast	14.74	11.86	11.24	50.89	40.96	50.74

Table 5.5.6.2 Comparison of forecast sources: Standardised Differences of Assessment measures for Probability Forecasts associated with Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Rates) (In this Table, the base forecast is "Most Likely" with either zero or 100% error)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Mean Absolute Error (mm/hr)	*(certain)*						
	Prob. Median	1.53	2.92	0.13	0.19	1.36	0.65
	$Const_{4mm/hr/hr}$	0.53	1.14	1.61	-2.76	-1.68	-1.30
	$Const_{30mm/hr}$	-0.82	0.77	-0.59	-4.09	-3.65	-2.41
Continuous Brier Score (mm/hr)	*(100% error)*						
	Prob. Median	1.18	2.75	1.17	0.21	1.01	0.24
	$Const_{4mm/hr/hr}$	-0.10	0.96	1.68	-3.06	-2.17	-2.16
	$Const_{30mm/hr}$	-1.34	0.31	-0.34	-3.66	-3.48	-3.31
	(operational)						
	Prob. Forecast	0.27	1.28	0.72	-1.57	-0.99	-1.61

Table 5.5.6.3 Comparison of forecast sources: Standardised Differences of Assessment measures for Probability Forecasts associated with Heavy Rainfall Warning forecasts in the Thames Region. (Rainfall Rates) (In this Table, the base forecast is "Prob. Forecast": the operational probability forecast)

Assessment Measure	Forecast Source	Raingauge Ground-Truth			Radar Ground-Truth		
		Area of Thames Region			Area of Thames Region		
		NE	SE	W	NE	SE	W
Continuous Brier Score (mm/hr)	*(100% error)*						
	Most Likely	-0.27	-1.28	-0.72	1.57	0.99	1.61
	Prob. Median	2.08	2.28	1.34	3.36	2.56	3.15
	$Const_{4mm/hr/hr}$	-0.22	0.70	1.25	-2.55	-1.87	-1.69
	$Const_{30mm/hr}$	-1.52	-0.14	-0.59	-2.63	-2.65	-2.75

5.5.7 Summary

Section 5.5 has described the results obtained for a case study concerning Heavy Rainfall Warning forecasts of rainfall in the Environment Agency's Thames Region. The targets of the operational rainfall forecasts in this case are of two types: the largest rainfall accumulation at any site and the largest rainfall rate at any site. These Warnings are issued separately for the three Areas of the Thames Region according to certain agreed criteria for when warnings should be issued. Unfortunately, the contents of the Warnings do not include a statement of why a Warning has been issued. The forecasts of rainfall amounts and rates relate to specific time-periods within which a rainfall event is forecast to occur: the lengths of these time-periods vary between Warnings.

The assessments described here have been performed on a forecast-by-forecast basis for a total of 18 warnings for the Northeast and Southeast Areas and 13 for the Western Area.

Two sources of ground-truth have been considered in the case of rainfall-rates: a raingauge network and the Nimrod-Quality Controlled radar product. The differences in spatial resolution provided by these products means that the corresponding ground-truths for spatial maximum rainfall amount differ markedly, with the radar source usually providing higher spatial maximum rates than the raingauges.

As for the case-study of Evening Update forecasts summarised in Section 5.4.7, two sources of ground-truth have been considered in the case of rainfall-rates: a raingauge network and the Nimrod-Quality Controlled radar product. The conclusions are much the same as earlier (*q.v.*), in that the forecasts are much better attuned to the ground-truth derived from raingauges than they are to the radar-derived ground-truth. For the purposes of this report, the apparent disparity between the forecasts and outcomes when the radar-based ground-truth is used suggests that the raingauge-based ground-truth should be used for any conclusions.

Results have been presented for a large range of measures of forecast performance, and the results have included comparisons of the operational forecasts against one type of simple forecast (constant-valued forecasts). In the case of forecasts of the maximum rainfall amount during the forecast period, the operational forecasts appear to perform better than these simple forecasts. The operational forecasts of the maximum rainfall rate appear to perform somewhat less well than those for the maximum rainfall totals, at least when the R^2-type of performance statistics are considered. The operational forecasts of the maximum rainfall rate do not seem to perform better than simple forecasts, and may actually be worse. However, the small dataset here than for rainfall totals means that this conclusion is not strongly supported.

Besides providing straightforward single-value forecasts, the Heavy Rainfall Warnings for this case study provide forecasts in the form of probability tables for the outcome that might occur. The analysis here has included an assessment of these probability forecasts. The results found suggest that these operational probability forecasts do not really perform better than a simple alternative probability forecast derived from the single-value stated as the main forecast (see Section 5.5.1).

The results shown for this case-study have included values for the standardised difference statistic which was described in Section 4.3. This statistic is designed for use in assessing whether there is enough evidence to support a conclusion that one forecast source is better than another, given that a direct comparison is subject to sampling uncertainty. The utility of this type of statistic has been successfully illustrated.

5.6 Summary

Section 5 has presented the results of analyses considering the performance of forecasts of rainfall for the three types of forecasts received by the Environment Agency. In most instances it is difficult to draw strong conclusions about forecast-performance because of the relatively small amount of data available for the comparisons. However, the main aims of this project have been achieved, namely

(i) the testing of a wide range of measures of forecast performance;
(ii) comparison of the performance of different potential sources of forecasts;
(iii) consideration of different sources of ground-truth for rainfall information.

These three aims are used to structure a summary of Section 5 below.

Assessing a range of measures of forecast performance

Results for a wide range of measures of forecast performance have been presented, with the widest range being used in the case-studies involving Evening Update forecasts and Heavy Rainfall Warnings. Each type of measure of performance is targeted at a slightly different aspect of how forecasts and outcomes differ, and thus, as might be expected, there is no single performance measure that stands out as the obvious single choice.

Of the direct measures of quantitative forecast performance, those based on errors in the logarithm of rainfall amounts and rates have the attraction of giving an intuitively reasonable balance between the importance of forecast errors for differing rainfall amounts. However, there are a number of different ways of overcoming the difficulties associated with this type of performance measure (arising from the treatment of zero rainfalls) each involving a somewhat arbitrary choice of adjustment parameter. We recommend that some consideration needs to be given to further variants of such logarithm-of-rainfall measures and also that an examination is needed of ways of improving the interpretation of these performance measure where they are used to provide implied bounds on what the outcome might be given the value of the forecast.

Some indirect measures of forecast performance have been considered. These are derived from a categorization of rainfall amounts according to whether of not the forecast and outcome amounts are above or below a selected threshold. This type of measure is well-established in the forecasting literature. These measures include several which have a readily-understood meaning in terms of being estimates of long-run proportions involving “success” or “failure” of the forecasts: however, at least two of these measures are needed to provide an adequate picture of how well the forecasts are doing (and this is just for a single choice of threshold, whereas several thresholds

would typically be considered at once, leading to further performance measures being evaluated). Some alternative measures have been defined which combine some of the basic measures in an attempt to provide a single overall measure of performance. The analysis here has found that using such measures can be problematic in achieving even the initial step of establishing a well-defined value for the measure when sample sizes are small, or when thresholds are rarely exceeded. This problem of ill-defined values is closely associated with the problem of estimation based on small samples: there seems to be a strong possibility that improved measures of performance of categorical forecasts can be achieved given some detailed investigations at a fairly basic level, by bringing-in statistical concepts rather than just probabilistic ones.

Results have been given for a third type of performance measure. This type is initially targeted specifically to be able to treat forecasts which are expressed as probability distributions describing the range of outcomes that may occur (as distinct from the more standard case of a single-valued forecast). Such measures can be applied to single-valued forecasts if some implied range of likely-forecast-error is used to construct a notional probability forecast. The case-studies here have shown that reasonable results are produced and that the measure has the potential to be used to compare the usefulness of forecasts from different sources each of which have associated (possibly incorrect) statements of likely accuracy.

The results of the analyses have emphasised the potential benefits of having access to measures of the accuracy with which the performance measures are determined by the sample sizes. Such assessments of accuracy are readily available for certain of the performance measures, but there would need to be some further development of similar assessments for the other performance measures, and it is recommended that this be undertaken. Assessments of the accuracy of performance measures are not usually quoted, partly because they provide an incomplete picture of the uncertainty remaining when comparing the performance of different forecast sources. This problem arises from the statistical dependence of the performance measures when applied to the different sources for the same events, which needs to be accounted for. The results presented here have not included direct measures of the accuracy of the performance measures, for reasons of both time and space. However, we have included results for a comparison of forecast sources which does take account of the statistical dependence between performance measures as well as the uncertainty in their estimation. This is seen to be a valuable form of analysis. It is unfortunate that it is not available for all types of performance measure and we recommend that research be undertaken to provide a way of achieving similar comparisons for an extended range of cases.

Performance of different potential sources of forecasts

The case-studies of the Daily Weather Forecasts have provided examples where the actual forecasts issued by the Met Office have been compared with Mesoscale Model and Nimrod forecasts. The standardised difference method of comparing forecasts has been demonstrated to be useful in determining whether there is enough evidence that one forecast has performed better than another. The results indicated that there were only some occasions when this was the case for the small sample sizes used in this assessment.

The case-studies involving the Evening Update forecasts and the Heavy Rainfall Warnings have been based on rather larger sample-sizes than those for the Daily Weather Forecasts and have included comparisons with a range of alternative forecasts sources, but these have been of a somewhat artificial character. The comparison of the Met Office forecasts against persistence forecasts and constant-valued forecasts are useful in showing that the Met Office forecasts do have some skill, but the results indicate only a minor improvement over persistence forecasts. This should not be interpreted as criticism of the Met Office forecasts but rather as indicating that the target quantities are very difficult to predict.

Both the Evening Update and the Heavy Rainfall Warning case-studies have included a comparison of the main operational forecast with a closely related forecast derived from the probability forecast. While this is not strictly a candidate source of forecasts for operational use, it has allowed example analyses to be made of the comparison between two similar forecast sources.

Each of the Evening Update and the Heavy Rainfall Warning case-studies have included a probability forecast as part of the operational product. No operational competitor of this type was available for comparison, but the study has included some simple alternative probability forecasts constructed by adding an uncertainty-range to the single-valued forecasts. The results of this analysis indicated that the operational probability forecasts do not perform better than such simple probability forecasts.

Different sources of ground-truth for rainfall information

The case studies have used two main source of ground-truth data. These have been the networks of recording raingauges operated by the Environment Agency and the Met Office, and the Nimrod Quality Controlled weather-radar product. The recording raingauge data were quality-controlled by cross-comparison within the network, but not compared with, or adjusted to match, the daily-read raingauge network.

The result of the comparison of the forecasts against these two sources of ground-truth were as follows. In the cases of the Evening Update forecasts and the Heavy Rainfall Warnings, the target quantities are spatial maxima and one would expect rather different maxima to be identified by the network of gauges and by the radar. It is clear from the results found that the operational forecasts are rather better attuned to the outcomes derived from the raingauge network than they are to the radar-derived values. It should be emphasised that this is a preliminary finding based on a period of record where the Nimrod product is of uncertain status in terms of it stage of development and the extent of availability of real-time raingauge information supporting its correction procedures.

In the case of the Daily Weather Forecasts comparison of both ground truth sources and methods of averaging was carried out. For Thames Region it was found that the "Typical" rainfall quantity was often an overestimate of the rainfall amount as measured by all the alternative forms of ground truth, and so it is difficult to recommend raingauge or radar as a better source of ground truth. Similarly it is difficult to recommend a single method of spatial averaging in forming the ground truth for this quantity, although a reduce set of truths derived from radar and

raingauges can be recommended. The findings concerning ground truth for the "Max" rainfall quantity were similar to those found for the Evening Updates and Heavy Rainfall Warnings. For Northeast and Northwest regions all ground truths used gave similar results. The mean raingauge ground truth was used for simplicity, although the alternative truths of radar or raingauge areal averages also seemed reasonable. However since only very small samples were used in the Northeast and Northwest regions if is difficult to recommend a single ground truth as suitable for all occasions.

6. SUMMARY AND CONCLUSIONS

6.1 Format and Content of Forecasts

Section 3 of this report includes a detailed review of the format and contents of the rainfall forecasts received by the Environment Agency for the regions served by the Met Office London Weather Centre. It also includes some discussion of the formats received by other Regions of the Agency. The review extends to considering how formats might be improved in the longer term.

The formats and styles of forecasts do change over time, and the ones reviewed here relate principally to July-November 2002. There is an intention jointly between the Environment Agency and the Met Office that the formats for forecasts used across all Regions should be brought into line. Proposals for such unified formats have used the existing formats of the London Weather Centre as a starting point. The current versions of the contents of these forecasts do have some serious shortcomings, even in terms of meeting the needs of the small collection of Regions for which they are presently used. Section 3.4 gives a few selected recommendations for changes, and some of the main points are summarised here and in Section 6.2.

At present, the London Weather Centre issues 3 types of rainfall forecasts: Daily Weather Forecasts, Evening Updates and Heavy Rainfall Warnings. This report suggests that the present style and content of Evening Updates, which for some reason has been made similar to that used for Heavy Rainfall Warnings, should be replaced entirely by a shortened version of whatever is eventually used for Daily Weather Forecasts (that is, exactly as used for the Daily Weather Forecasts but with the time-horizon covered truncated at the end of the day following the day of issue). Regions not served by the London Weather Centre do not presently receive Evening Updates: the fact that they occasionally receive corrections to earlier Daily Weather Forecasts when conditions change unexpectedly does not seem to be an adequate replacement for the regular Evening Update service.

At present certain minor difficulties over the exact interpretation of the forecasts arise because of the use of local time (i.e. GMT or BST) within the forecasts rather than using only the standard GMT/UTC convention. We suggest that the forecasts should be based on the GMT/UTC convention, but with an adequate prompt to the possible need to convert to local time being given in heading information, for example by stating the issue time in both GMT and local-time forms. We believe that most Environment Agency staff receiving rainfall forecasts will be working with data from other systems, in particular telemetry and flow forecasting systems, which will also be using GMT/UTC as the principal basis.

The formats and contents of Heavy Rainfall Warnings should be revised to reflect the sets of criteria that state when the service should provide Warnings. At present the actual contents of the Warnings (from the London Weather Centre and some others) are rather separated from these criteria and there is essentially just a forecast of a rainfall amount over a certain period, the criteria for which is unexplained. The sets of criteria for warnings are carefully thought-out by the Environment Agency in terms of rainfall conditions that are meaningful to them for catchment-response considerations. It seems important that the Warnings received should provide meaningful information relating to such conditions.

6.2 Target Forecast Quantities

The cases studies have involved a number of different quantities as targets of the forecasts. The target-quantities relate to values defined for pre-specified sub-areas of an Environment Agency Region. Each type of target is defined in a slightly different way in terms of values at individual sites within each sub-area: the three types of targets of main concern here are:

(i) the mean (average) value over the sub-area;
(ii) the maximum value over the sub-area;
(iii) the "typical" value for the sub-area.

Many Regions receive forecasts for sub-area averages and for maximum values, but the forecasts requested for those Regions served by the London Weather Centre presently include "typical" values for the sub-areas.

The "typical" value is notionally the value at a typical site in the sub-area being considered. In practice, this is an ill-defined concept and the way we have attempted to represent the calculation of this quantity from actual data may not agree with anyone else's interpretation. However different people will have different interpretations. The term "mode" or "modal value" has sometimes been used as describing what is required, and sometimes it has been implied that it is the "mode" across those sites which receive a positive amount of rainfall.

The maximum value within a sub-area is a better-defined quantity, but it suffers from the problem that the value being targeted by the forecast is affected by the resolution of the ground-truth against which the forecasts are tested. Thus, values of spatial-maxima of rainfall obtained from a network of raingauges will be larger if the network is made more dense. While rainfall derived from weather-radar may appear to overcome this problem, a similar problem does arise with radar-derived rainfalls if possible changes in spatial resolution are considered. In practice, some tuning of forecasts of spatial maxima against a particular source of ground-truth data is likely to have occurred, if only in an informal way.

The values at individual sites from which the target-quantities are derived are, in most cases, rainfall totals over a given time-period. In the case of the forecasts provided by the London Weather Centre, the target quantities include some derived from the maximum rainfall-rate within a given time-period: the maximum of the site-values then gives the highest rainfall intensity experienced anywhere in the sub-area. The specification of maximum-rainfall-rate as a target quantity is again problematic, since the value, even for a single site, will depend radically on the time-step at which the underlying rates are notionally defined. In theory, rates might be defined at a one milli-second or a one second time-step and the maxima of these set of rates would be expected to be substantially higher than rates derived from 5 or 15 minute rainfall totals. Once again, in practice, the forecasts of maximum rainfall rates are likely to have been tuned against a source of ground-truth data in which a particular way of specifying the rainfall rates has been determined. It seems that the forecasts of rainfall rates agree better with values derived from 15-minute raingauge totals than they do with the (quasi-instantaneous) rainfall rates estimated by weather radar.

In the case-studies of the Daily Weather Forecasts, the target quantities have been based on rainfall totals over interval-lengths of 6, 12 and 24 hours, which together cover a total time-period of 5 days. For the Thames Region case study the target quantities were "typical" and maximum values for each of three sub-areas.

For the case-study of the Evening Updates, the target quantities are maximum rainfall totals and maximum rates for a single time-period 18 hours in length, for each of 3 sub-areas.

For the case-study of the Heavy Rainfall Warnings, where forecasts for the Thames Region are used, the target quantities are again maximum rainfall totals and maximum rates for a single time-period. But, in this case, the interval lengths are set by the forecasters in response to what they foresee as being the extent of the rainfall event.

6.3 Ground Truth

The case-studies have made use of two sources of information about rainfall to provide the ground-truth against which the forecasts can be compared. These are, firstly, the network of recording raingauges for which records are available either directly as 15-minute rainfall totals or as times-of-tips. Secondly we have used the Nimrod Quality-Controlled product to provide radar-derived rainfalls.

The results from the case-studies indicate that when ground-truths are used based on raingauge-networks or on the Nimrod product, results are comparable only for those cases where the target of the forecast is a spatial average rainfall. Sections 5.4.2 and 5.5.2 contain examples of the forecasts and outcomes (according to raingauge network and radar ground-truths) in cases where the target of the forecast is the maximum rainfall amount or rate within an area. These clearly show that the raingauge- and radar-derived ground truths for spatial maxima are radically different, and that the forecasts currently provided by the Met Office are a much better match for the raingauge-derived ground-truth than for the radar ground-truth. Some caution is required here because it is not clear that the service provided by Nimrod Quality-Controlled product has stabilised sufficiently to warrant firm conclusions. It is important that a confirmation is obtained from users of the forecasts that the forecast-target should be the largest pixel-value in a radar image derived from this product before the Met Office adjust their forecasting procedures to match this target. In principle, the problem identified here relates to the specification of the target value to match user-requirements and not directly to finding the best estimate of the rainfall over an area. This is a different topic and one which is addressed next.

Section 4.2 has discussed the various sources of data that might be available in the longer term to provide estimates of the rainfall for use as ground truth when assessing the performance of rainfall forecasts. In principle, the best source for such ground-truth data is likely to be a combination of raingauge and radar data. Two main sources of such data are either presently available or planned to be available from the Met Office on a (near-)real-time basis: the Nimrod Quality-Controlled and “Merged” products. These, being locally archived, would be available to the Environment Agency Regions for rainfall-forecast performance monitoring. Other raingauge-radar combination procedures are available. Section 4.2 has identified, as a point for further investigation, the question of whether the quality-control of raingauge-data (and of radar data, to a lesser extent) used in such real-time products is sufficiently good to lead to reliable estimates of the rainfall fields. If there are no such problems, then the archived datasets could be used for the routine monitoring of rainfall forecasts. However, it would still be best to undertake a more thorough re-estimation of the ground-truth for more definitive investigations or in cases of comparing forecasts where the differences are likely to be small. Such re-estimation could well bring in data from a more-

extensive collection of raingauges than used for the real-time procedures. If the combined raingauge and radar products are found to be badly affected by problems relating to the quality-control of raingauge data, then re-processing of the corrected (and possibly extended) datasets is indicated. Since the raingauge-radar combination procedures in operational use are presently undertaken by the Met Office, the question arises as to whether they could provide such a re-processing service.

The question of quality-control of raingauge data is clearly an important one, and one particular aspect of this needs to be made explicit. It is commonly accepted that ordinary telemetering and recording raingauges cannot be relied on to measure precipitation during periods of snowfall, nor to provide an indication of snowmelt. Section 4.2 considered the possibility of using information from daily-read raingauges, where procedures for dealing with accumulated snowfall are in place, but these data would not provide values for the same 24-hour periods as used in Daily Weather Forecasts. The arrangements for snow-observers to record information about lying snow vary across the country, but these observers are unlikely to provide a sufficient spatial coverage and the information will usually only be recorded daily. In practice, any periods affected by falling, lying or melting snow will need to be omitted from quantitative assessments of forecasts because there will be insufficient information about precipitation over short-periods on an area-wide basis.

Accessibility of raingauge data for assessment is another issue. The Rainfall Collaboration Project is providing shared access to raingauge data from networks operated separately by the Met Office and by the Agency, but which conform to certain selection and quality control criteria. The HARP project is providing the Agency with a data archiving system encompassing raingauge data. These developing infrastructures need to be taken into account when considering quality control and use of raingauge data for forecast assessment purposes. The interfaces to systems for forecast assessment, including both automated and ones initially based on manual data entry, require careful consideration.

6.4 Forecast Sources

The present report has shown how different forecasts for the same target quantities can be compared. The analyses in Sections 5.4 and 5.5 have derived slightly artificial forecast values from the information contained in the overall operational forecasts to provide alternatives to actual forecast values for those occasions. This has provided a means of presenting examples of the comparison of forecasts in cases where no such operational alternatives were available, without resorting to totally unrealistic forecasts such as “always zero”. The results found suggest that data from considerably more forecast occasions than has been available for the present study are needed in order to make a definitive distinction between the performances of closely similar forecasts.

In the case of the rainfall content of Daily Weather Forecasts, it has been possible to undertake a comparison of the forecasts issued with corresponding forecasts obtained from the Mesoscale Model: these analyses are described in Section 5.3. In principle, the forecasts actually issued are a blend of the information in these model-forecasts with Met Office forecasters’ knowledge and experience, and their interpretation of available synoptic analyses and radar/satellite information. Thus this particular comparison is a test of whether the Met Office forecasters are able to improve the underlying automatically-produced forecasts. The Mesoscale Model forecasts used here for areal rainfalls were derived for this project from the

raw gridded model forecast results. This is a point that still needs clarifying, but it seems that the routine procedures used by the Met Office to construct forecasts start with results which are automatically derived from gridded model results, but from the Global Model rather than the Mesoscale Model. The results presented in Section 5.3 suggest that the operational forecasts distributed in the Daily Weather Forecasts have properties which are rather different from those of the Mesoscale Model forecasts. In particular the operational forecasts appear to be typically larger than the eventual outcomes while the Mesoscale Model forecasts are effectively unbiased. Unfortunately the datasets available for analysis are too small to allow a definitive conclusion about the comparative forecast performance to be made, but there is weak evidence that the Mesoscale Model forecasts are the better ones. It should be noted that the case-study events span a period from January to November 2002, whilst on 26 August 2002 the mesoscale/global models changed to employ the "New Dynamics" of the Unified Model including use of smoother orography. No consideration of this change has been taken in the comparisons.

The case-studies of the Daily Weather Forecasts have allowed the comparisons to include forecasts derived from weather-radar forecasts of rainfall. This is only possible for the first 6-hour forecast within each day as the time-horizon for the radar forecasts is only 6 hours. There are some difficulties in interpreting the comparison of forecasts here, since the operational forecasts in the DWFs are typically based on model-runs and observations available at midnight, whereas the radar-based forecasts would be based on observations up to 6 am. However, the Met Office forecasters do have access to rainfall information later than midnight, and could use it to modify model-based forecasts. The result for 6-hour forecasts derived from the Nimrod product indicate a fairly good performance, but the amount of data used for these comparisons has not been sufficient to justify a conclusion that the radar-forecasts for the first period of the day are definitely better than the corresponding forecasts in the operational Daily Weather Forecasts. The limited time-horizon of Nimrod rainfall forecasts means that this source of forecasts is not appropriate for use as a competitor to the operational Heavy Rainfall Warnings and Evening Updates, where the time-intervals covered are rather longer.

The analyses presented in Section 5 have included not only "ordinary" forecasts, where the forecast consists of a single "best-estimate" value, but also "probability forecasts", where the forecast consists of a set of values that together specify the probability that the eventual outcome will be within any given range. The case studies showed that these operational probability forecasts do not compare favourably to simple equivalents derived by adding a fixed allowance of uncertainty to the "best-estimate" forecast. However, it can be argued that procedures for probability forecasting are only just now being developed within the Met Office and that it is therefore too soon to abandon this type of service. The extent of detail about possible outcomes provided by a probability forecast of the type studied here is thought to provide a considerable improvement over the simple two-category alternative, which simply states a probability that the rainfall amount will exceed a given threshold. Unfortunately, reasonably complete probability forecasts take up a lot of space within a forecast compared to the simpler two-category forecast.

The assessments of forecast performance presented in Section 5 have included certain types of unrealistic forecasts for comparison. Their primary purpose was to check whether the assessment procedure provided a sensible performance ranking of alternative forecast methods. These forecasts have been of two types. For the first type, known as "persistence

forecasts", a simple forecast is constructed by setting the forecast-value according to the value observed for the "ground-truth" for a time-period before the beginning of the forecast period. The second type of forecast, "constant-valued forecasts", sets the forecast-value in some fixed way, possibly related to interval-length, but one which does use any observations of recent rainfall. The results found that the persistence forecasts can have a reasonable performance for the first forecast period compared to the operational forecasts, possibly because they make use of more up-to-date observed rainfall information than the DWF and Evening Update forecasts, which are issued somewhat in advance of the start of the first forecast period.

6.5 Performance Measures

Section 2 has outlined a large number of different measures of forecast performance and has outlined some of the advantages and disadvantages of these. Many of the measures have been used within the case studies presented in Section 5 and extensive tables of results have been presented.

The results in Section 5 have exemplified various obvious problems with some of the performance measures. In particular, some of the measures related to categorical forecasts may not be fully defined when applied to small sets of data. In the case of performance measures related to errors in the logarithm of rainfall, there are somewhat arbitrary decisions made in the treatment of zero-rainfalls that can have an important effect on the result calculated. Performance measures that are based on errors in log-rainfall are often used when assessing the behaviour of forecasting schemes. In principle, the choice between using errors in rainfall or errors in log-rainfall relates to the question of how the size of forecast errors relate to the amounts of rainfall being forecasted and observed, with the use of log-rainfalls being suggested where the sizes of errors are typically proportional to the amounts of rainfall. While it is intuitively appealing to think that the proportional-errors behaviour might hold, it is not necessarily true. The scatter plots of observed against forecast data shown in Section 5 do not have this type of behaviour for any of the different forecast datasets, and hence the use of errors in log-rainfall is not indicated.

Section 4.3 has discussed the question of assessing how well a measure of forecast performance is determined by the set of data used within an analysis, and it has also discussed the question of assessing whether a given set of data provides enough evidence to make a clear distinction between the performances of different forecast-sources. These are important questions that are often not addressed in simple lists of forecast-performance results. The case-studies presented in Section 5 have included results relating to distinguishing between forecast sources. These results suggest that, while in most cases there is enough evidence to say that the operational forecasts really are better than simple constant-valued and persistence forecasts, the datasets are too small to allow a clear distinction between similar operational forecasts. The resources for this phase of the project have not allowed for the evaluation of the accuracy with which the individual performance measures are determined.

The results reported in Section 5 are presented in a way that treats the Areas within a Region separately, rather than combining results for the forecasts across Areas into a single answer. Since there is usually no reason to expect radically different performance for the forecasts across the Areas, the variation in results across the Areas provides some indication of how well the performance measure is determined, which can be useful in the absence of a more

formal evaluation. Similarly, the results for the Daily Weather Forecasts have been presented in a way that shows the variation in performance as a function of the forecast lead-time. It would be expected that, if evaluated for a large dataset, the performance should vary smoothly with lead-time: thus again the non-smooth variation with lead-time gives an indication of how well the performance measures are determined.

From a mathematical point of view, the choice of performance measure should be at least partly based on the interpretation given to the forecast-values in terms of what they represent in the context of uncertainty about what the outcome will be. Unfortunately, this is not particularly helpful in the present instance where no firm intention is stated, where interpretations may vary from Region to Region and from person to person and where use of mathematically ambiguous phrases like "most likely value" is prevalent. The relationship of performance measure to the interpretation of the forecasts can be stated as follows. Consider an evaluation based on a large set of forecast occasions containing only meteorologically similar cases. Then, in those instances where the interpretation of what would be counted as good forecasts is that the average value of the forecasts should be the same as the average of the outcomes, this implies that the forecasts are essentially the expected value of the probability distribution of the unknown outcome given the meteorological information on which the forecast is based. Performance measures such as the root mean square error have a natural association with forecasts that have this type of expected-value interpretation. In contrast, the interpretation of what would be counted as good forecasts might be that half of the outcomes should be above the forecast value and half below. This implies that the forecasts are essentially the median value of the probability distribution of the unknown outcome given the meteorological information on which the forecast is based. The mean absolute error has a natural association with forecasts which have this type of median-value interpretation.

A simple example can be constructed to illustrate the properties of the mean absolute error (*mae*) and root mean square error (*rmse*). Suppose for simplicity that there are 100 forecast occasions, and that all outcomes are zero. Suppose two forecast sources give the following forecasts for these 100 occasions.

Forecast 1 : 99 values where forecast is 0, and 1 forecast of 100
Forecast 2 : 99 values where forecast is 1, and 1 forecast of 99.

Then the following performance measures apply:

Forecast 1: *mae* = 1.00; *rmse* = 10
Forecast 2: *mae* = 1.99; *rmse* = 9.95.

This illustrates that, in moving from Forecast 1 to Forecast 2, the *rmse* criterion is dominated by the (relatively unimportant) improvement to the single large forecast error while the large number of occasions where the small errors are made worse are effectively ignored. In contrast, the *mae* criterion treats all changes to forecast errors on an equitable basis.

In practice, the performance measures listed in Section 2 are each targeted at different aspects of forecast performance and so it is good to have a number available. Where the requirement is for a performance measure that gives a meaningful value that can be readily interpreted, the mean absolute error seems to be a good choice. Firstly, it gives a "typical size of error" in the same units as the rainfall data (ie. mm or mm h^{-1}). Further it is easily understood and it is fairly stable, in that it gives a reasonable assessment of performance across a number of

forecast occasions, without being swamped by a single occasion on which a forecast is particularly bad. The Critical Success Index and False Alarm Rate form a pair of measures that are widely used, for assessing forecast performance, in situations related to correctly forecasting whether a given threshold of rainfall amount will be exceeded. The main problem with these is choosing a relevant threshold: it is common practice to use a range of thresholds. The Odds Ratio measure has been designed for use where the purpose is to track changes of forecast performance over time. This particular purpose is not of prime importance here and the measure suffers badly in being undefined or extremely variable when applied in cases where the sample size is small. It seems possible that an alternative measure of a similar type and intention can be developed which would overcome this deficiency, and this might be a valuable contribution to the science of comparing forecasts. However, the above-mentioned pair of category-based performance measures are likely to provide an adequate summary of performance. Where probability forecasts are being assessed, the Continuous Brier Score (the only one found for this context) has the property that it is rather similar to Mean Absolute Error, but adjusted to take into account whether the implied range of possible values is too narrow or too wide. It again provides a measure of accuracy that is on the same scale as the rainfall quantities.

Table 6.5.1 provides a summary of the performance measures that have been selected as most important. These include the already mentioned measures: mean absolute error, CSI and FAR skill scores, and the Continuous Brier Score for use with probability forecasts. The table also includes measures of bias since, operationally, there is interest in whether forecasts are consistently overestimating or not, in addition to their "typical error size". The median error is preferred to the normal bias (mean error) as it is more robust to outliers. In the same way that the mean error and median error complement each other, the root mean square error is included to complement the mean absolute error. The *rmse* is more sensitive to a few large errors, which can highlight problem forecasts (or problem ground-truths) but is not as robust an indicator of "typical error size". The R^2 Efficiency statistic is also included as providing a dimensionless and relative variant of the *rmse* measure. As a measure of bias for categorical forecasts, the Bias Ratio is included for similar reasons as for the two continuous measures of bias. The Probability of Detection is included to complement the CSI and FAR statistics. In addition, the Odds Ratio and Likelihood Ratios have been included in the present selection of more useful performance measures. These are seen as important for assessing forecast system performance independent of the natural variability of weather over time and place, but they have a lesser value to the forecast practitioner interested in the overall accuracy of the forecast. The ordinary Brier Score has been included to cover applications where probability forecasts cannot be converted into distributions covering the whole range of possible values: for example where the forecast states a probability that a given single threshold will be exceeded.

Appendix B provides a readily accessible guide to the selected performance measures with a simple example of their application in assessing forecasts from the Heavy Rainfall Warning product. It also provides a succinct guide to forecast assessment, including the procedures for comparing forecasts from different sources and using different ground-truths. Appendix C is included for readers requiring a deeper understanding, in probability terms, of the Contingency Table for Categorical Skill Scores and the Likelihood Ratios and Odds Ratio that derive from it.

Table 6.5.1 Selected Performance Measures

(a) Continuous Measures

Assessment Measure	Formula
Basic quantities	y_i is the observed value of rainfall for sample i (i=1,2..., n). $\hat{y}_i$ is the forecast value of rainfall for sample i.
Bias (mean error)	$bias = n^{-1}\sum(y_i - \hat{y}_i)$
Median error	50% point of errors
Mean absolute error	$mae = n^{-1}\sum\lvert y_i - \hat{y}_i\rvert$
Root mean square error	$rmse = \sqrt{n^{-1}\sum(y_i - \hat{y}_i)^2}$
R^2 (Efficiency)	$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$ $\bar{y} = n^{-1}\sum y_i$ is the sample mean of the observations.

(b) Categorical Measures (Skill Scores)

Categorical Skill Scores

Contingency table:

Event Forecast	Event Observed		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	a+b+c+d

Critical Success Index (Threat Score)	$CSI = \frac{a}{a+b+c}$
False Alarm Rate	$FAR = \frac{b}{a+b}$
Probability of Detection (Hit Rate for observed 'yes')	$POD = \frac{a}{a+c}$
Bias Ratio	$B = \frac{a+b}{a+c}$

Table 6.5.1 cont' Selected Performance Measures

(b) Categorical Measures (Skill Scores) (cont')

Likelihood Ratio	LR_1 is the Likelihood Ratio for correct forecast of "below X" $LR_1 = \frac{d(a+c)}{c(b+d)}$ LR_2 is the Likelihood Ratio for correct forecast of "above X" $LR_2 = \frac{a(b+d)}{b(a+c)}$
Odds Ratio	$\theta = \frac{ad}{bc} = LR_1 LR_2$

(c) Skill Scores for Probability Forecasts

Categorical	
Brier Score	$BS = n^{-1}\sum (Y_i - \hat{Y}_i)^2$ Y_i indicator of event $y_i \le x$ in the observed sample, equal to 1 if event $y_i \le x$ does occur, 0 if not $\hat{Y}_i$ probability of event $y_i \le x$ occurring , as stated in the probability forecast, value in the range 0 to 1 Here y_i is the observed value of sample *i* (*i*=1,2..., *n*), and *x* is a threshold value defining the categories of event-occurrence or non-occurrence.
Continuous	
Continuous Brier Score	$BS = n^{-1}\sum \int (Y_i(x) - \hat{Y}_i(x))^2 dx$ Y_i (x) indicator of event $y_i \le x$ in the observed sample, equal to 1 if event $y_i \le x$ does occur, 0 if not $\hat{Y}_i$ (x) probability of event $y_i \le x$ occurring, as stated in the probability forecast, value in the range 0 to 1 Here y_i is the observed value of sample *i* (*i*=1,2..., *n*), and *x* is a variable threshold value covering all possible values of rainfall amount or rate.

6.6 Forecast Assessment Procedures

The type of procedures for assessing rainfall forecasts discussed in this report can be characterised as working on a forecast-by-forecast basis. There is an alternative basis for forecast assessment which might be said to operate on a rolling basis. This approach is outlined briefly in this subsection, after considering the advantages and disadvantages of the approach used in this report.

The forecast-by-forecast approach is a relatively simple procedure, in that it works by taking each of the forecasts issued and compares the forecast quantities with their corresponding outcomes. This a reasonable basis for assessing the Daily Weather Forecasts and Evening Updates, since these are issued regularly, but is arguably a poor basis for assessing Heavy Rainfall Warnings, since it does not allow account to be taken of occasions when a Warning is not issued and when a high rainfall amount did occur. The forecast-by-forecast approach has another problem in that it does not readily allow any account to be taken of the difference between the times at which a forecast is issued and the start of the first forecast-period. In the case of Heavy Rainfall Warnings, forecasts may actually be issued several hours after the start of the time-period for which a Warning is issued. In fact, this feature may mean that, in practice, Heavy Rainfall Warnings are always issued in relation to high-rainfall events, so that the question of events not being counted in the assessment may not arise. However, this does reveal the somewhat ambiguous nature of the forecast-by-forecast assessment procedure when applied to Heavy Rainfall Warnings.

The main advantage of the forecast-by-forecast approach to forecast assessment is that the requirements for ground-truth data are relatively modest, which means that it is suited to computer implementation where data values are entered manually. An assessment of Heavy Rainfall Warnings on a rolling basis requires substantially more data for use as ground-truth and hence needs to be reserved for more fully automated assessment procedures.

A rolling-basis forecast assessment procedure would be structured so as to consider a regularly-spaced set of base-times at which a comparison of forecasts and eventual outcomes is made. Only forecasts received prior to a given base-time would be considered as eligible for inclusion in the comparison. The comparison would be between the outcome for a given time-period and the most most-recently arrived forecast for that time-period. In principle this approach would allow a relevant comparison to be made between forecasts-products where one might be delivered relatively early while the one delivered somewhat closer to the target time-period might be more accurate. Such an approach would be generally applicable and it would theoretically be possible to consider situations where DWFs, Evening Updates and Heavy Rainfall Warnings are assessed together as a single rainfall-information service, as well as separately.

The main difficulty in treating Heavy Rainfall Warnings within a rolling-basis assessment procedure arises from the fact that, as discussed in Sections 4.1 and 6.1, the rainfall forecasts contained in the Warnings do not relate to specific time-periods and the Warnings (from London Weather Centre) do not state against which criteria the Warning has been issued. The rolling-basis forecast assessment procedure would require that, in both cases where a Warning has been issued or has not been issued, some meaningful quantitative interpretation can be made of the Warning in terms of the rainfall to be expected within a given time-period. Non-issuance of a Warning, where there are clear criteria for when warnings are to be issued,

would correspond to a forecast that rainfall will be below a given bound. This may initially lead to the use of performance measures for categorical forecasts being applied to a rolling-basis analysis of HRWs. However, if the quantitative information contained in Heavy Rainfall Warnings is to be more fully assessed, it may be necessary to develop performance criteria more directly suited to mixed categorical-quantitative forecasts.

6.7 Performance of Forecasts

Section 5 has presented a detailed analysis of forecast performance for the various types of forecast. In general it seems that the performance of the operational forecasts is poor. This may be most clearly seen in the scatter plots of observed against forecast rainfall, where no clear relationship between the two is evident. The analytical results do show that the performance of the operational forecasts is better than naïve forecasts (persistence and constant-valued forecasts), and better than would be obtained by chance. Thus the apparent poor performance could well be due to the difficulty of the forecasting task. The present operational forecasts are not clearly outperformed by either of the two potential competitors for operational use: forecasts derived directly from the Mesoscale Model, and radar-based forecasts (which would be available for a lead-time of only 6-hours). However, there are indications that these might actually be better, but it would take a larger dataset to demonstrate this clearly. Note also that the dataset encompassed events in the period January to November 2002 whilst the "New Dynamics" of the Unified Model were used as the basis of operational NWP forecasts from 26 August 2002.

6.8 Conclusions

(a) Format and Content of Forecasts

(1) The contents and format of the different types of forecasts should be considered and specified jointly, so that consistent definitions and terminology are used.

(2) All time-periods within the forecasts should be specified directly on the GMT/UTC scale. Actual issue-times for forecasts should appear explicitly within each forecast and these issue-times should be given on both local and GMT/UTC time-scales.

(3) The present style of Evening Updates from the London Weather Centre should be entirely replaced by a shortened form of a revised Daily Weather Forecast where, for Evening Updates, this would be restricted in time-coverage to finish at the end of the next day.

(4) The style of Heavy Rainfall Warnings should be revised to state why the warning was issued in relation to the agreed criteria for when Warnings should be issued. This is particularly important for Regions where more than one criterion is used.

(5) Consideration should be given to a new separate service for warnings of high rainfall 2-5 days in advance to be broadly equivalent to the present service received by the Agency's Northeast Region. It seems unlikely that a new style for Heavy Rainfall Warnings, similar to

the service presently received by other Regions, would be appropriate for such long lead-times.

(6) Automated production and receipt of Mesoscale Model rainfall fields should be considered. An operational trial involving the Study's assessment procedure should be initiated.

Note*: A number of changes to the formats of the operational forecasts have been made subsequent to the case-study periods considered in this report, and a brief note of these changes has been included at the end of Section 3. This includes item (4) above.*

(b) Target Forecast Quantities

(7) The principal target quantities for forecasts should be the average (arithmetic mean) rainfall within an area and the maximum rainfall, where these would both be total rainfalls over prescribed periods. Where rainfall rates need to be targeted, careful consideration is needed of the relevant space and time-scales for these. We suggest that the smallest realistic scaling would be to define rates in terms of averages over 2×2 km^2 radar-pixels and over 15-minute or one-hour time-periods.

(8) The analysis of forecasts of spatial maxima has suggested that these are presently tuned to provide reasonable values for the maximum across a network of raingauges, but do not agree with the maxima derived from weather radar.

(9) The present style of Heavy Rainfall Warnings gives a forecast rainfall total for a variable-length time-period. A revision of the style to more closely match the criteria for issuing warnings would be to take fixed-length periods of say 6 and 12 hours and, for each of these period-lengths, to state those times when the running total of that length is expected to exceed the agreed threshold.

(c) Ground Truth

(10) The most suitable ground-truth for use in assessing forecast performance may well differ between the different possible purposes of the assessment. Routine assessment procedures require that the ground-truth be derived from data that are readily available. When comparing the performance of different sources of forecast, effort should be expended in creating the best possible version of ground truth utilising relevant data sources and subjecting them to careful quality control procedures. Periods affected by snowfall will usually need to be excluded from assessments. The infrastructures provided by the Rainfall Collaboration Project and HARP in providing access to ground-truth data need to be considered in the design of computer systems for rainfall forecast assessment.

(11) A ground-truth derived only from local networks of raingauges is likely to be adequate only when the forecast target is a spatial average rainfall. In principle, information from a merged radar-raingauge product would be preferred.

(12) Ground-truths for cases where the forecast target is a spatial maximum rainfall should be based on a product that employs both weather radar and raingauge sources, such as the merged product under development at the Met Office.

(13) An assessment is needed of whether the weather-radar products distributed for real-time use are suitable for post-event analyses. Potential areas of concern relate to the short-term occurrence of the usual range of radar-problems that can be difficult to reliably correct in real-time, and to the quality control of the raingauge information, which may again be difficult to handle fully on a real-time basis.

(14) There is a potential need for a full re-analysis of rainfall fields based on radar and raingauge data which have both been quality-controlled on a post-event basis, and the implications of this need to be thought through. The need for such a re-analysis partly depends on how adequate the real-time analysis is found to be. If no problems are found, then the Agency could well use the real-time data for routine forecast performance assessments, but would otherwise need to have access to re-processed datasets.

(d) Forecast Sources

(15) Procedures have been described which allow a statistical comparison to be made between the performances of forecasts from different sources. In particular, these procedures assess whether there is enough evidence to decide that one source is better than another. Further work is needed to extend the range of forecast measures for which such procedures are available to encompass categorical skill scores.

(16) A comparison has been made between the operational forecasts issued by the Met Office to the Environment Agency and forecasts derived, via a simple areal averaging procedure, from the Mesoscale Model rainfalls. Unfortunately, the datasets available have been too short to allow a definitive conclusion to be made; they also spanned a period during which the "New Dynamics" of the Unified Model were introduced as the basis of the operational NWP model forecasts. Based on the case study events analysed, the present operational forecasts have the feature of over-estimating rainfalls, while the Mesoscale Model forecasts do not.

(17) A comparison has also been made of the operational forecasts for the initial 6-hour time-period with those available from the Nimrod radar product. Again, the datasets available have been too short to allow a definitive conclusion to be made. Certainly neither is dramatically better than the other.

(18) Simple forecasts are important in showing that forecasts have at least some skill. Two types have been considered in this report: persistence forecasts and constant-valued forecasts. It is suggested that forecast assessment procedures will always need to include comparisons of operational forecasts against such simple forecasts.

(e) Performance Measures

(19) A wide range of performance measures has been reviewed. This report contains a summary of the perceived advantages and disadvantages of the performance measures. These

measures are all potentially useful in that they measure different aspects of forecast performance.

(20) It is suggested that a basic set of forecast performance measures for use within a simple assessment tool should be:

Continuous variable:

Bias (mean error)	over- or under-estimation
Median error	over- or under-estimation
Mean absolute error	typical size of error
Root mean square error	typical size of error
R^2 (Efficiency)	size of error relative to a simple forecast

Categorical variable:

Critical Success Index	balanced measure of forecast success
False Alarm Rate	emphasises events incorrectly forecasted
Probability of Detection	emphasises events correctly forecasted
Bias Ratio	too many or too few events forecasted
Likelihood Ratios	measure of information provided by having forecast service, separately for events and non-events
Odds Ratio	overall measure of information provided by having forecast service

Where probability forecasts are analysed:

Brier Score	error in probability terms
Continuous Brier Score	balanced measure of location and spread of forecasts relative to outcome

Visual:

Scatter plots of outcome against forecast.	Visual appreciation

(f) Forecast Assessment Procedures

(21) The type of procedures for assessing rainfall forecasts used in this report work on a forecast-by-forecast basis. This seems to be most appropriate for a simple assessment tool where data-entry is made manually.

(22) An alternative, rolling-basis, procedure for assessing forecasts can be envisaged and this would be more appropriate than the forecast-by-forecast basis for the case of Heavy Rainfall Warnings since account can be taken of instances where a Warning has not been issued.

REFERENCES

Selected Bibliography on Forecast Performance Measures

Applequist, S., Gahrs, G.E., Pfeffer, R.L. and Niu, X.-F. 2002. Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Weather and Forecasting*, **17**, 783-799.

Buizza, R. 2001. Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Monthly Weather Review*, **129**, 2329-2345.

Doswell III, C.A., Davies-Jones, R. and Keller, D.L. 1990. On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting*, **5**, 576-585.

Göber, M., Wilson, C.A., Milton, S.F. and Stephenson, D.B. 2003. Fairplay in the verification of operational quantitative precipitation forecasts. *J. of Hydrology*, 27pp, submitted.

Golding, B.W. 1998. Nimrod: A system for generating automated very short range forecasts. *Meteorol. Appl.*, **5**, 1-16.

Jollife, I.T. and Stephenson, D. B. 2003. *Forecast Verification: A Practitioner's Guide in the Atmospheric Science*. Wiley.

Marzban, C. 1998. Scalar measures of performance in rare-event situations. *Weather and Forecasting*, **13**, 753-763.

Murphy, A.H. 1991. Probabilities, Odds, and forecasts of rare events. *Weather and Forecasting*, **6**, 302-307.

Murphy, A.H. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281-293.

Schaefer, J.T. 1990. The Critical Success Index as an indicator of warning skill. *Weather and Forecasting*, **5**, 570-575.

Stanski, H.R, Wilson, L.J. and Burrows, W.R. 1995?, *Survey of common verification methods in meteorology*. World Weather Watch Technical Report No. 8, WMO/TD. No. 358, World Meteorological Organisation, 114pp.

Stephenson, D.B 2000. Use of the "Odds Ratio" for diagnosing forecast skill. *Weather and Forecasting*, **15**, 221-232.

Van Den Dool, H.M. and Toth, Z. 1991. Why do forecasts for "near normal" often fail? *Weather and Forecasting*, **6**, 76-85.

Wilks, D.S. 1995. *Statistical methods in the atmospheric sciences. An Introduction*. (Chapter 7, Forecast verification, 233-282), Academic Press.

Appendix A Calculation of the Continuous Brier Score

The Continuous Brier Score for a set of probability forecasts and the corresponding outcomes is

$$B = n^{-1}\sum_{i=1}^{n} B_i(y_i),$$

where $B_i(y_i)$ is the contribution from the i'th forecast occasion, when the observed outcome is y_i. The quantity $B_i(a)$ is determined by the probability forecast, that was distributed for the i'th forecast occasion, as a function of a, where a is used to denote a value for the outcome that might have been observed: the actual outcome is $a = y_i$.

For the purpose of this project, the probability forecasts are in the form of a simple probability table which has been converted to a full probability distribution by assuming that the uncertainty is uniformly distributed over intervals separating the points in the probability table. Here, a single forecast occasion is considered and the symbol $P(x)$ is used to denote the probability distribution corresponding to the probability forecast for that occasion. The forecast tables are interpreted in such a way that $P(x)$ consists of a discrete component of size p_0 at x_0, where usually $x_0 = 0$, together with a set of disjoint uniform components, each of total probability p_j, on intervals defined by the possibly irregular sequence $x_j = x_{j-1} + \delta_j$ $(j = 1,2,...)$.

The probability p_j and the interval length δ_j apply to the interval (x_{j-1}, x_j) for $j = 1, 2, \ldots$, and the probabilities, including that at zero, sum to one $\left(\sum_{j=0}^{\infty} p_j = 1\right)$.

For the given value of a, define $j = j(a)$ as the smallest integer for which for $a \le x_j$. Then the contribution to the overall Continuous Brier Score from this forecast occasion is given by

$$B(a) = (a - x_0) + 2C_{j(a)}(a) + \sum_{l=1}^{\infty} E_l - 2\sum_{l=1}^{j(a)} D_l.$$

where

$$\begin{aligned} C_j(a) &= (x_0 - a) && (j = 0), \\ &= (x_j - a)\{U_j + \tfrac{1}{2} p_j (x_j - a)/\delta_j\} && (j \ge 1), \end{aligned}$$

and

$$\begin{aligned} D_j &= \delta_j (U_j + \tfrac{1}{2} p_j) && (j \ge 1), \\ E_j &= \delta_j (U_j^2 + p_j U_j + \tfrac{1}{3} p_j^2) && (j \ge 1), \\ U_k &= \sum_{j=k+1}^{\infty} p_j = 1 - \sum_{j=0}^{k} p_j, && (k \ge 1). \end{aligned}$$

Figure A.1 shows two examples of the Continuous Brier Score function, *B(a)*. In the first case, the probability forecast is concentrated on a fairly narrow range, while in the second the probability forecast is spread over a much wider range. The plots for these cases are shown on the same scales so that a direct comparison can be made. Recall that the Continuous Brier Score measures the error or discrepancy between a probability distribution (the forecast) and

an observation. In a sense the observation should be treated as fixed, with the Continuous Brier Score measuring how well alternative probability forecasts accord with the observation. In the examples in Figure A.1, the score for an observation close to the centre of the narrow probability forecast receives a much lower score than does an observation close to the centre of the more wide-spread forecast: this illustrates that the Continuous Brier Score penalises the wider probability distribution in this instance for being comparatively wide. An observation of 120mm would receive a high score for the narrow probability distribution, but a lower one for the wider distribution. In this case the narrow distribution scores badly both because it is centred on values which did not occur and because the distribution is narrow. An observation of 0mm would receive a low score for the narrow probability distribution, and a high one for the wider distribution. In this case the value is on the edge of both distributions but the narrow distribution places its weight closer to the outcome than does the wider distribution. In these examples, the wider distribution is actually bimodal. This has little evident effect on $B(a)$ in this case. In general, the Continuous Brier Score function is always convex-downwards and there is always a single minimum (although the function may be constant over an interval).

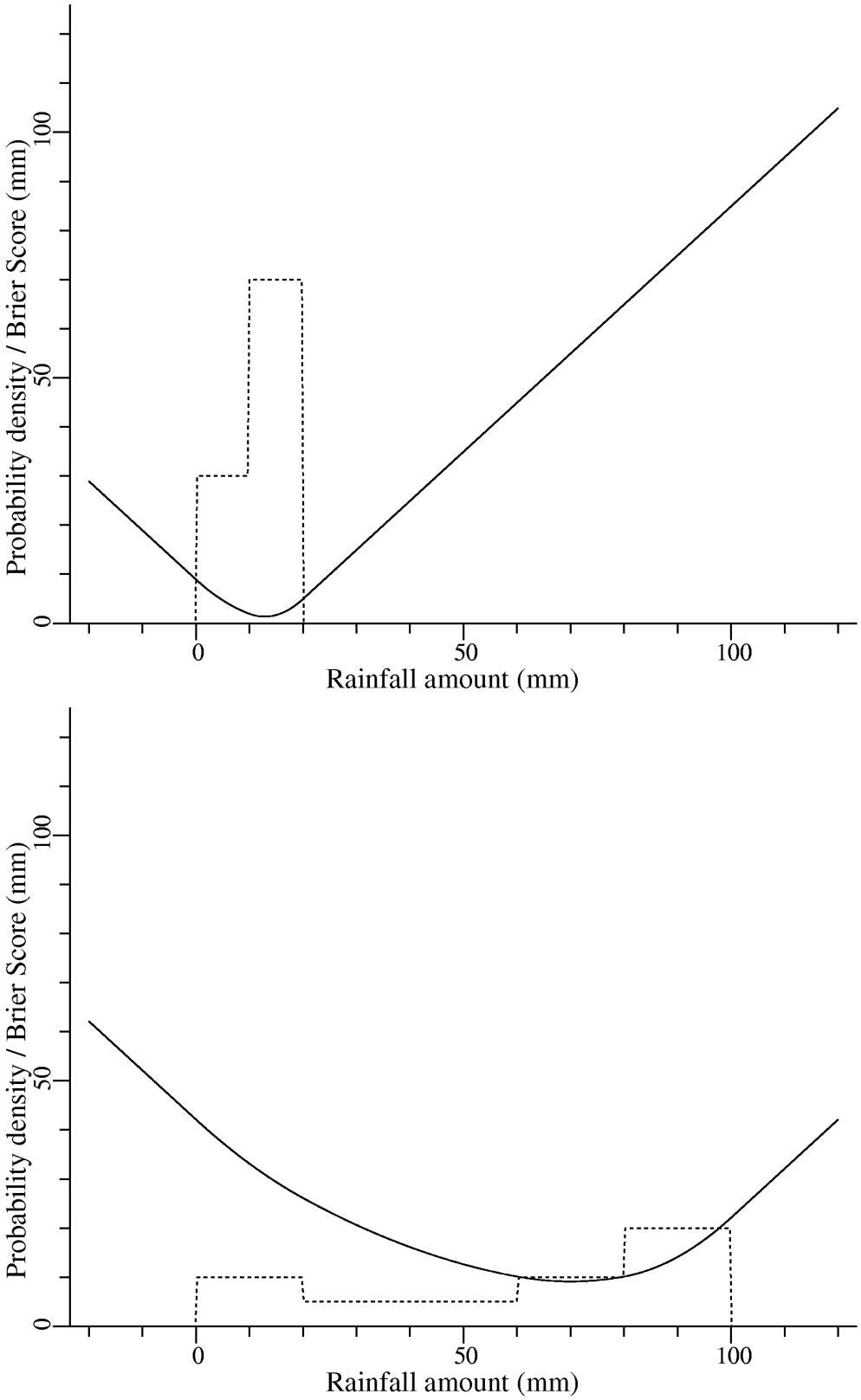


Figure A.1 **Examples of the Continuous Brier Score as a function of the observed rainfall amount, for cases where the probability forecast (dashed) is relatively concentrated or wide-spread.**

Appendix B A Guide to Performance Measures and Assessment using the Heavy Rainfall Warning Assessment Tool

B.1 Introduction

This Appendix provides a guide to the performance measures and assessment procedure recommended in this report. It is written with specific reference to the Heavy Rainfall Warning product as it features as Section 6 of the User Guide to Heavy Rainfall Warning Assessment Tool, a PC software product with manual data-entry developed under the Study. It is included here as an Appendix, as it provides an easily accessible guide to the main features of the assessment procedure and the performance measures involved, including a simple example of their use.

An overview of the performance measures, recommended by this Study for use in the assessment of the rainfall forecast products, is presented in Table B.1. Section B.2 presents a basic guide to assessment, covering the choice of ground truth, comparative forecasts and routine for assessment. Section B.3 contains a guide to a selection of the most commonly used performance measures with examples for each. Section B.4 contains an overview of the performance measures not covered in Section B.3. Finally, Section B.5 provides a guide to making comparisons of different forecast sources and ground truths.

Table B.1 Overview of performance measures

Continuous variable:	
Bias (mean error)	over- or under-estimation
Median error	over- or under-estimation
Mean absolute error	typical size of error
Root mean square error	typical size of error
R^2 (Efficiency)	size of error relative to a simple forecast
Categorical variable:	
Critical Success Index	balanced measure of forecast success
False Alarm Rate	emphasises events incorrectly forecasted
Probability of Detection	emphasises events correctly forecasted
Bias Ratio	too many or too few events forecasted
Likelihood Ratios	measure of information provided by having forecast service, separately for events and non-events
Odds Ratio	overall measure of information provided by having forecast service
Where probability forecasts are analysed:	
Brier Score	error in probability terms
Continuous Brier Score	balanced measure of location and spread of forecasts relative to outcome

B.2 Guide to Assessment of Heavy Rainfall Warnings

The Tool's assessment procedures for Heavy Rainfall Warnings are aimed at answering the basic question:

> *What is the typical size of error in rainfall forecasts, or rate of success in forecasting high rainfalls?*

The Tool can also be used to monitor changes in forecast performance over time.

Use of the HRW Assessment Tool requires that the target quantities for the forecasts can be properly identified, so that suitable ground-truths can be identified, evaluated by the user and entered into the Tool. The target quantity of a forecast may be an average for an area, or a maximum within an area. Similarly, the target may be a total rainfall for a time-period (or equivalently an average rate over a time-period), or a maximum rate within a time period. Where the target quantities relate to "maxima", difficulties arise over defining this concept in a way that gives a meaningful value, taking into account the effect of using data of differing spatial or temporal resolution. There can also be questions over defining a "typical" rainfall value if the average value (specifically the mean value) is not quite what is required. For use of the Assessment Tool, values of ground truth should be prepared according to whatever interpretation of "ground-truth" is acceptable, bearing in mind any discussions between the Met Office and the Environment Agency concerning what targets of the forecast should be.

The interpretation of individual Heavy Rainfall Warnings for use within the Tool can be problematic where the Warnings do not have a fixed structure. The Tool requires matched sets of forecast amounts or rates and corresponding ground-truths, which, in principle, requires that the Warnings be interpreted as providing quantitative forecasts for specific areas and specific time-periods.

The HRW Assessment Tool is designed for the situation where entry of all data required will be accomplished manually. This has affected the choice of assessment procedures, leading to the adoption of a forecast-by-forecast based assessment procedure. This means that the Tool's assessment procedure measures how well the forecasts of rainfall contained in the Warnings perform in matching the eventual outcomes. An assessment of the success or failure for the issuing of Warnings against the criteria for when they should be issued cannot be undertaken with the Tool because this would require a different approach and would required substantially more data than can be handled conveniently using manual data-entry.

The Assessment Tool provides facilities for comparing forecasts from a number of sources using a number of different versions of ground-truth. Typically the operational Heavy Rainfall Warnings should be compared against simple forecasts that are constructed according to simple rules, thus providing a performance baseline. One simple forecast is to always forecast a fixed amount, say 20mm, while values for another can be constructed by forecasting an amount which is proportional to the length of the event (as contained in the Heavy Rainfall Warning).

Ground-truths for rainfall may derive from two different types of information: raingauge networks and weather radar. In principle, merging of information from raingauge networks

and weather radar should be the best source of ground truth but this is problematic at present. Spatial averages may be adequately estimated by raingauges alone, by (adjusted) radar alone, or by use of a fully merged product. Theoretically, spatial maxima would be best estimated using radar data because of the higher spatial resolution, but experience has shown that forecasts of maxima may be better matched to the maxima obtained from a raingauge network.

The Nimrod "merged" product does not yet exist to provide experience on which advice can be based and the Nimrod "Quality Controlled" product is still undergoing changes and development. The suitability of locally-archived Nimrod data for post-event analyses has not yet been assessed: there may be a need for post-event quality control of Nimrod data. Similar problems arise for other radar-raingauge products constructed for real-time use.

If ground-truth is obtained from a raingauge network, then these data should be adequately quality-controlled. Procedures for simple inter-gauge comparisons are required. The possible effects of incorrect raingauge data having been used within merged radar-raingauge products would need to be considered.

B.3 Guide to Performance Measures, Part 1

B.3.1 Example

For the purposes of illustrating the performance measures, the example below has been used.

A set of 5 forecast Heavy Rainfall Warnings of the Spatial Maximum Accumulation (mm) are to be assessed for Northeast Region South Pennines Area. Radar data (Nimrod QC 2km) have been selected to provide the ground-truth. Radar has the potential to provide a better spatial maximum rainfall estimate than use of data from a raingauge network. (Note that this may not be the case in practice due to problems with radar rainfall estimation.) The values concerned are tabulated below.

Start Time/Date	Forecast HRW	Ground truth Radar
15:00 29-Jul-02	30	189.88
15:00 30-Jul-02	60	102.78
03:00 1-Aug-02	60	46.47
08:00 4-Aug-02	15	34.09
06:00 9-Aug-02	30	51.88

B.3.2 Guide to Notation

y_i is the observed (ground-truth) value of rainfall for sample i (i=1,2..., n).

$\hat{y}_i$ ***is the forecast value of rainfall for sample i.***

Summation operator, $\sum$

$$\sum y_i \equiv \sum_{i=1}^{n} y_i = y_1 + y_2 + ... + y_n$$

Example: Mean of (n=5) radar observations

$$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{1}{5}(y_1 + y_2 + y_3 + y_4 + y_5)$$

= 0.2 (189.88+102.78+46.47+34.09+51.88)= 85.02 mm

B.3.3 Continuous Performance Measures

Bias (mean error)

Mean of the rainfall forecast errors.

$$bias = n^{-1}\sum(y_i - \hat{y}_i)$$

Use: Indicates over-estimation (negative) or under-estimation (positive) of rainfall forecast.

Example: Bias of Heavy Rainfall Warning forecasts

Forecast errors, $y_i - \hat{y}_i$, are: 159.88, 42.78, -13.53, 19.09, 21.88.
$bias = n^{-1}\sum(y_i - \hat{y}_i)$ = 0.2 (159.88+42.78-13.53+19.09+21.88) = 46.02 mm
This indicates forecast underestimation by 46 mm.

Median error bias

Median of the rainfall forecast errors.

50% point of errors

Use: Indicates over-estimation (negative) or under-estimation (positive) of rainfall forecast.

Example: Median error of Heavy Rainfall Warning forecasts

Forecast errors, $y_i - \hat{y}_i$, ranked in order of size are: 159.88, 42.78, 21.88, 19.09, -13.53
The Median Error is given by the 50% point of errors, which is 21.88 mm.

This indicates forecast underestimation by 22 mm, whilst mean error bias is 46 mm.
The median error as a measure of bias is *more robust to outliers* than the mean error, giving a more typical bias value in this case.

Mean absolute error

Mean of the absolute values of the rainfall forecast errors.

$$mae = n^{-1}\sum|y_i - \hat{y}_i|$$

Use: Typical size of rainfall forecast error.

Example: Mean absolute error of Heavy Rainfall Warning forecasts

Absolute value of forecast errors, $|y_i - \hat{y}_i|$, are: 159.88, 42.78, 13.53, 19.09, 21.88.
$mae = n^{-1}\sum|y_i - \hat{y}_i|$ = 0.2 (159.88+42.78+13.53+19.09+21.88) = 51.432 mm.

Root mean square error

Square root of the mean of the squared rainfall forecast errors.

$$rmse = \sqrt{n^{-1}\sum(y_i - \hat{y}_i)^2}$$

Use: Typical size of rainfall forecast error.

Example: Root Mean Square Error of Heavy Rainfall Warning forecasts

Square of forecast errors, $(y_i - \hat{y}_i)^2$, are: 25562, 1830, 183, 364, 479.
$rmse = \sqrt{n^{-1}\sum(y_i - \hat{y}_i)^2}$ = √{0.2 (25562+1830+183+364+479)} = 75.39 mm.

Compare the *typical size of error* given by *mae* of 51.432 mm with the *rmse* value of 75.39 mm. The *rmse* is more sensitive to outliers, as seen in this example where the value calculated is inflated by taking the square of the single large error value of 159.88. The *rmse* is arguably less *typical* than the estimate provided by the *mae* estimator.

R^2 Efficiency

Proportion of variance in observations accounted for by forecast.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

$\bar{y} = n^{-1}\sum y_i$ is the sample mean of the observations

Use: Size of error relative to a simple (sample mean) forecast (dimensionless)

Example: R^2 Efficiency of Heavy Rainfall Warning forecasts

The sample mean of the radar observations has previously been calculated as $\bar{y}$ = 85.02 mm.
The observed deviations from the mean, $y_i - \bar{y}$, are: 104.86, 17.76, -38.55, -50.93,-33.14.
The sum of squares of these deviations is
$\sum(y_i - \bar{y})^2$ = 10996+315+1486+2594+1098 = 16489.
The term $\sum(y_i - \hat{y}_i)^2$ is obtained from the *rmse* value of 75.39 previously calculated, by squaring and multiplying by 5 to give 28418. Then:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - (28418/16489) = -0.72.$$

This indicates that the set of HRW forecasts are 72% worse than a simple constant forecast equal to the mean of the radar observations (note this simple forecast cannot be realised in practice as the set of radar observations are not known at the time of forecast construction).

B.3.4 Categorical Skill Scores

Contingency table:

Event Forecast	Event Observed		Total
	Yes	No	
Yes	*a* hit	*b* false alarm	*a+b*
No	*c* miss	*d* correct rejection	*c+d*
Total	*a+c*	*b+d*	*n=a+b+c+d*

An Event is defined as an exceedence of a rainfall threshold value.

a, *b*, *c* and *d* are the number of entries in each Event category for *n* rainfall forecasts and their corresponding observations.

Example: The performance of the Heavy Rainfall Warning forecast in warning of rainfall events in excess of 49 mm is to be assessed.

The Contingency Table for this rainfall event threshold and tabulated set of rainfall values is readily calculated as:

Event Forecast	Event Observed		Total
	Yes	No	
Yes	*1* hit	*1* false alarm	*2*
No	*2* miss	*1* correct rejection	*3*
Total	*3*	*2*	*5*

This indicates that there are 3 observed events exceeding the 49 mm threshold, of which 1 is correctly forecast (a hit) and 2 are missed, whilst there is 1 false alarm and 1 correct rejection of an event.

Critical Success Index (Threat Score), CSI

Number correct (hits) divided by number forecast and/or observed (the threat: a+b+c)

$$CSI = \frac{a}{a+b+c}$$

Use: Balanced measure of forecast success.

Example: CSI performance of the Heavy Rainfall Warning forecast in warning of rainfall events in excess of 49 mm is to be assessed. Using the Contingency Table entries gives:

$$CSI = \frac{a}{a+b+c} = \frac{1}{1+1+2} = 0.25 .$$

False Alarm Rate, FAR

Proportion of forecast events that fail to materialise.

$$FAR = \frac{b}{a+b}$$

Use: Emphasises events incorrectly forecasted.

Example: FAR performance of the Heavy Rainfall Warning forecast in warning of rainfall events in excess of 49 mm is to be assessed. Using the Contingency Table entries gives:

$$FAR = \frac{b}{a+b} = \frac{1}{1+1} = 0.5 \cdot$$

Probability of Detection (Hit Rate for observed 'yes'), POD

Proportion of occasions when an event does occur that are forecasted to experience the event.

$$POD = \frac{a}{a+c}$$

Use: Emphasises events correctly forecasted.

Example: POD performance of the Heavy Rainfall Warning forecast in warning of rainfall events in excess of 49 mm is to be assessed. Using the Contingency Table entries gives:

$$POD = \frac{a}{a+c} = \frac{1}{1+2} = 0.33 \cdot$$

Bias Ratio, B

Ratio of "yes" forecasts with "yes" observations.

$$B = \frac{a+b}{a+c}$$

Use: Indicates too many (greater than 1) or too few events (less than 1) forecasted.

Example: Bias Ratio performance of the Heavy Rainfall Warning forecast in warning of rainfall events in excess of 49 mm is to be assessed. Using the Contingency Table entries gives:

$$B = \frac{a+b}{a+c} = \frac{1+1}{1+2} = 0.67 .$$

In summary, the Skill Scores are: CSI=0.25, FAR=0.5, POD=0.33 and B=0.67. Thus there is a tendency to under-forecast, with a Bias Ratio B less than 1 and a low False Alarm Rate.

B.4 Guide to Performance Measures, Part 2

B.4.1 Relative Categorical Skill Scores

Likelihood Ratios, LR_1 and LR_2

LR_2 is the Likelihood Ratio for correct forecast of an event.

$$LR_2 = \frac{a(b+d)}{b(a+c)}$$

The chance of forecasting that an event will occur when that event does happen is LR_2 of the chance of forecasting the event will occur when it actually does not.

LR_1 is the Likelihood Ratio for correct forecast of a non-event.

$$LR_1 = \frac{d(a+c)}{c(b+d)}$$

The chance of forecasting that an event will not occur when that event does not happen is LR_1 of the chance of forecasting the event will not occur when it actually does happen.

A good forecast service will have Likelihood Ratios greater than 1.

Use: Measure of information provided by having forecast service, separately for events and non-events

Example: Likelihood Ratio performance of the Heavy Rainfall Warning forecast in warning of rainfall events in excess of 49 mm is to be assessed. Using the Contingency Table entries gives:

$$LR_2 = \frac{a(b+d)}{b(a+c)} = \frac{1(1+1)}{1(1+2)} = 0.67\,.$$

The chance of forecasting that the event will occur when the event does happen is *2/3* of the chance of forecasting the event will occur when it actually does not.

$$LR_1 = \frac{d(a+c)}{c(b+d)} = \frac{1(1+2)}{2(1+1)} = 0.75\,.$$

The chance of forecasting that the event will not occur when the event does not happen is *3/4* of the chance of forecasting the event will not occur when it actually does happen.

Odds Ratio, θ

Compares the conditional odds of making a good forecast (a hit) to those of a bad forecast (a false detection).

The *odds* (or risk) Ω of an event is the ratio of the probability *p* of it occurring to it not occurring, *1-p,* and so $\Omega = p/1-p$.

$$\theta = \frac{ad}{bc} = LR_1 LR_2 .$$

Odds of an observed event being correctly forecast is the *Odds Ratio* times the odds of a no-event being forecast as an event.

Use: Overall measure of information provided by having forecast service.

Example: Odds Ratio performance of the Heavy Rainfall Warning forecast in warning of rainfall events in excess of 49 mm is to be assessed. Using the Contingency Table entries gives:

$$\theta = \frac{ad}{bc} = \frac{1\times 1}{1\times 2} = 0.5 .$$

Alternatively, the product of the two Likelihood Ratios, 0.67 times 0.75, gives the same result.

Thus, the odds of an observed event being correctly forecast is half the odds of a no-event being forecast as an event. A good forecast service has an Odds Ratio greater than 1.

Appendix 3 provides a probability interpretation of the Relative Categorical Skill Scores that provides a path for gaining a deeper understanding of these skill scores.

B.4.2 Skill Scores for Probability Forecasts

Brier Score (Categorical)

Mean square probability error.

$$BS = n^{-1}\sum (Y_i - \hat{Y}_i)^2$$

Y_i indicator of event $y_i \le x$ in the observed sample,
equal to 1 if event $y_i \le x$ does occur, 0 if not

$\hat{Y}_i$ probability of event $y_i \le x$ occurring ,
as stated in the probability forecast, value in the range 0 to 1

Here y_i is the observed value of sample *i* (*i*=1,2..., *n*), and *x* is a threshold value defining the categories of event-occurrence or non-occurrence

Use: Typical size of error in probability terms.

Brier Score (Continuous)

Integrated mean square probability error.

$$BS = n^{-1} \sum \int (Y_i(x) - \hat{Y}_i(x))^2 dx$$

Y_i (x) indicator of event $y_i \le x$ in the observed sample,
equal to 1 if event $y_i \le x$ does occur, 0 if not

$\hat{Y}_i$ (x) probability of event $y_i \le x$ occurring,
as stated in the probability forecast, value in the range 0 to 1

Here y_i is the observed value of sample *i* (*i*=1,2..., *n*), and *x* is a variable threshold value covering all possible values of rainfall amount or rate.

Use: Balanced measure of location and spread of rainfall forecasts relative to outcome.

B.5 Guide to Making Comparisons with the HRW Assessment Tool

B.5.1 Guide to Comparing Forecast Sources

The value of simple forecasts for comparison against the operational forecasts was indicated in Section B.2. Good operational forecasts should out-perform simple forecasts. This leads to the question of comparing the performances of forecasts from different sources. When sample sizes are small, or when there is little difference in performance, any apparent difference may be due entirely to chance. The *standardised difference* of a performance measure for two forecast sources is used to indicate the extent of evidence that one source of forecasts is better than another.

The standardised difference is available for performance measures of the basic form

$$P = n^{-1} \sum g(outcome, forecast)$$

where *n* is the number of forecasts assessed and *g(.,.)* is some error function of the forecast and outcome (ground-truth) values (for example, the forecast error squared for *rmse*-type performance measures).

Then the difference in the performance measures for two sources is the average value of the differences

$$x_i = g(outcome, forecast^{(1)}) - g(outcome, forecast^{(2)}).$$

One forecast is better than another if the long-run average of the x_i's is different from zero. The evidence for or against one source being better than another can be quantified by the value of the standardised difference, t, where

$$t = \frac{\text{mean of } x_i\text{'s}}{\text{typical error of mean of } x_i\text{'s in estimating long - run average}}.$$

The standardised difference, *t*, is evaluated from the sample mean, $\bar{x}$, and sample variance, s^2, of the differences, x_i, as follows:

$$\bar{x} = n^{-1} \sum x_i$$

$$s^2 = (n-1)^{-1} \sum \{x_i - \bar{x}\}^2$$

$$t = \frac{\bar{x}}{\sqrt{n^{-1}s^2}}.$$

The standardised difference should be compared against the following suggested limits to assess whether there is reasonably strong evidence that one forecast source is better than another:

±2	if sample size n is large
±2.1	if sample size n =20
±2.5	if sample size n =10
±3.5	if sample size n =5.

B.5.2 Guide to Comparing Ground Truths

There are often several different ways in which ground-truth for rainfall quantities can be determined, particularly where the targets of rainfall forecasts is unclear. The HRW Assessment Tool can be used to make a comparison of ground-truths with the aim of assessing whether the forecasts are better matched to one version of ground-truth than another. The method for doing this is rather similar to comparing different sources of forecasts (Section 6.4.1), and a full account is not given here. Once again a standardised difference approach can be used for some types of performance measure.

The standardised difference is available for performance measures of the basic form

$$P = n^{-1}\sum g(outcome, forecast).$$

Then the difference in the performance measures for two gound-truths is the average value of the differences

$$x_i = g(outcome^{(1)}, forecast) - g(outcome^{(2)}, forecast).$$

Then the procedure for treating these x_i's is exactly the same as in Section B.5.1.

Appendix C Probability interpretation of Relative Categorical Skill Scores

Contingency table in probability form

Event Forecast	Event Observed		Total
	Yes	No	
Yes	$p(f,o)$	$p(f,\bar{o})$	*p(f)*
No	$p(\bar{f},o)$	$p(\bar{f},\bar{o})$	$p(\bar{f})$
Total	*p(o)*	$p(\bar{o})$	1

$p(f,o)$*:* joint probability of a hit being forecast and observed (a yes/yes event)
p(f): marginal probability for an event being forecasted
$p(f \mid o)$: *conditional probability* of a yes forecast given a yes observation
Overbar: signifies a no event eg. $\bar{f}$ indicates forecast is that an event will not occur

Likelihood Ratios, LR_1 and LR_2

LR_1 is the Likelihood Ratio for correct forecast of a non-event.

$$LR_1 = \frac{d(a+c)}{c(b+d)} = p(\bar{f} \mid \bar{o}) / p(\bar{f} \mid o)$$

LR_2 is the Likelihood Ratio for correct forecast of an event.

$$LR_2 = \frac{a(b+d)}{b(a+c)} = p(f \mid o) / p(f \mid \bar{o})$$

Use: Measure of information provided by having forecast service, separately for events and non-events

Odds Ratio, θ

Compares the conditional odds of making a good forecast (a hit) to those of a bad forecast (a false detection).

The *odds* (or risk) Ω of an event is the ratio of the probability *p* of it occurring to it not occurring, *1-p,* and so $\Omega = p/1-p$.

$$\theta = \frac{\Omega(f \mid o)}{\Omega(f \mid \bar{o})} = \frac{ad}{bc} = LR_1 LR_2 .$$

Use: Overall measure of information provided by having forecast service.