Defra/Environment Agency
Flood and Coastal Erosion Risk Management
R&D Programme

Performance measures for flood forecasting:

Review and recommendations

R&D Technical Report W5C-021/2b/TR

defra
Department for Environment
Food and Rural Affairs

ENVIRONMENT
AGENCY

Performance Measures for Flood Forecasting

Review and recommendations

R&D Technical Report W5C-021/2b/TR

Authors
Peter Hawkes, HR Wallingford
Chris Whitlow, EdenVale Modelling Services

Publishing organisation

ISBN: 1-8443-2401-X April 2005
Product code: SCHO0305BIVZ-E-E/-P

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

Statement of use
This report presents technical information and research findings from R&D Project W5C-021/2b (SC020076). The project provides recommendations for the definition and implementation of performance measures for the Environment Agency's flood forecasting capability. It will be of interest to those involved in operational real-time flood forecast modelling, and constitutes an R&D output from the Joint Defra / Environment Agency Flood Management R&D Programme. This report also constitutes HR Wallingford Report SR 663.

Keywords
Accuracy, flood forecasting, flood risk, flood warning, performances measures, reliability, timeliness.

Research contractor
This report was produced by HR Wallingford, Howbery Park, Wallingford, Oxon, OX10 8BA. The project manager was Peter Hawkes, pjh@hrwallingford.co.uk .

Environment Agency project manager
Andrew Grime, Weetwood, Elm House Farm, Saighton Lane, Saighton, Chester, CH3 6EN, andrew.grime@weetwood.net.

Contract Statement
This report describes work funded by the Environment Agency, as part of the Joint Defra / Environment Agency Flood Management R&D Programme, within the Flood Forecasting and Warning Theme Advisory Group, headed by Tony Andryszewski. The Environment Agency Project Codes were W5C-021/2b and SC020076. The HR Wallingford Job Number was CDS0828.

Further copies of this report are available from: The Environment Agency's National Customer Contact Centre by emailing enquiries@environment-agency.gov.uk or by telephoning 08708 506506. You may also download free from:www.defra.gov.uk/environ/fcd/research

SUMMARY

Flood forecasting in England and Wales is the responsibility of the Environment Agency, with separate operational forecasting rooms in each of the eight Environment Agency Regions. Flood forecasting is an essential component of an overall flood forecasting and warning service comprising:

- **Monitoring** of environmental variables that may lead to flooding, for example river flow, rainfall, tide, surge and waves, coupled with **Detection** of high values amongst the monitored variables
- **Forecasting** of possible flood incidents when one or more threshold levels are exceeded amongst the environmental variables
- **Warning**, in the sense of decision making on impending flood incidents when flooding risks are realisable
- **Dissemination** of warning messages to the public, professional partners and emergency services
- **Response** to the threat.

The purpose of the service is to organise effective action to reduce potential damage and injury. To achieve this, flood forecasting needs to be timely, accurate, reliable, cost-effective and well integrated into the overall service. In terms of day-to-day delivery, the main job of flood forecasters is to predict when certain pre-determined thresholds of certain flood risk variables at certain locations will be exceeded, with sufficient time for appropriate actions to be taken in response to the threat.

Fluvial flood forecasting is well developed in all eight Environment Agency regions, where outputs include peak water level, peak flow rate, time of peak and volume of flood. Except in the North West Region, coastal flood forecasting, including sea level, wave height and overtopping rate, is less well developed then fluvial. Over the coming few years, all regions will adopt the National Flood Forecasting System, which will bring consistency to the procedures and information used in both fluvial and coastal flood forecasting.

This report covers **performance measures**, which could be used to evaluate the effectiveness of individual sub-processes and/or the entire flood forecasting and warning service. In particular, this report focuses on **performance measures for flood forecasting**. Performance measures provide objective and consistent information on the performance of various models and procedures to Environment Agency staff involved in developing and improving flood forecasting services. The ultimate aim is to improve their effectiveness, with a clear focus to maintain or improve lead-time for the accurate, timely and reliable delivery of flood forecasting and warning services.

A wide array of performance measures can be used by developers and forecasters to help in identification of weak links in source data, modelling and procedures, to be prioritised for improvement. Quantitative mathematical measures include maximum error, mean error, bias, standard deviation and lead-time error. These are produced continuously as part of the forecasting process, and also at intervals to look at performance over a period of time. Simpler summary measures of potential flood events, such as skill scores and ‘contingency tables’, provide a comparison between

different implementations, and can be used to demonstrate value to funders and to non-specialists.

PART ONE of this report describes the wider context of flood forecasting, warning and response, within which performance measures would be applied. It reviews current practice in evaluation of flood forecasting services, and deficiencies, aspirations and ways forward to achieving a consistent system of performance evaluation.

PART TWO, comprising Chapter 4 and 5, gives the recommendations of this project. Chapter 4 begins by summarising the key points from PART ONE, then goes on to definitions of performance measures already planned to be implemented within NFFS, and additional performance measures that would be desirable additions to NFFS. These additional measures include some high level summary statistics, namely the contingency table (and associated skill scores) and a reliability measure. Chapter 4 concludes by discussing some of the practical issues associated with operation of these performance measures, especially in situations where measured data are unavailable.

Chapter 5 begins by summarising the alternative options for performance measures, and the relationship of those options to existing methods to be implemented within NFFS. It goes on to recommend one of those options, and to outline a programme for implementation and testing within the Environment Agency Regions. The recommendation is, initially, to accept all the performance measures already being developed within NFFS, plus several additional parameters. These include the R^2 correlation coefficient parameter already used within the Agency, a new timing error parameter incorporating a variable time lag between forecasts and measurements, a new reliability parameter and a contingency table. Part of the recommendation is to review this long list of performance measures after initial trials within the Environment Agency, to see if any are not used by forecasters and could be dropped. Chapter 5 concludes by outlining further R&D required to facilitate performance evaluation.

The Environment Agency will formulate a policy for implementation and continued development of the recommendations in this report, after completion of the project.

This document also constitutes HR Wallingford Report SR 663.

ACKNOWLEDGEMENTS

The Project Team consisted of Peter Hawkes, Chris Whitlow and Paul Sayers. The Project Board, consisting of five members, provided guidance and comment to the Project Team at regular intervals throughout the project. The members were:

Rahman Khatibi	Environment Agency (Head Office)
Andrew Grime	Weetwood
Ben Lukey	Environment Agency (North West Region)
Richard Cross	Environment Agency (Midlands Region)
Nigel Outhwaite	Environment Agency (Head Office)

GLOSSARY

Accuracy (of forecasts)
Refers to the agreement between forecast and subsequently measured or observed values and timings of variables, considered either continuously, or at peak or threshold crossing points, for different lead-times.

Alert (in context)
A term used by the Met Office, comparable to 'warning' used in the Environment Agency.

Contingency table
A small table summarising forecasting performance over a period of time, in terms of numbers of occurrences of measured and/or forecast exceedences of threshold levels, and the numbers of agreements and disagreements between the two.

Detection (of potential flood events)
The first informal signal, during monitoring of environmental variables, that a potential flood event may be coming. This may trigger an enhanced level of monitoring of environmental variables.

Dissemination (of flood warnings)
Referring to the distribution of flood warnings to Environment Agency staff, emergency services, the media and the public.

Flood forecasting (in context)
Forecasting of flood risk variables and pre-set threshold conditions based on those variables (excluding weather and ocean forecasting, and stopping short of warnings and responses).

Flood warning
Refers to the time (and the associated heightened activity and readiness) during which flood warnings are issued to the public.

Flood watch
Refers to a first level of public warnings being issued (and the associated heightened activity and readiness) when low impact flooding is possible, perhaps of farmland or minor roads,

Lead-time (before an event)
A modelling term referring to the length of time before the expected crossing of a threshold or the peak of a storm, at which model predictions (e.g. of threshold crossings) become available.

Level-to-level correlations
Refers to inferential estimation of river levels or sea levels, based on measurements at other nearby locations and empirical relationships between levels at the different locations.

Mitigation (of potential flood damage)
Refers to the actions that might be taken to reduce risk in times of potential flood events, and the reduction in injuries and damage to property that might be achieved as a result.

Monitoring (of environmental variables)
The routine assimilation of weather and offshore forecasts, measured variables, coupled with a background level of flood forecast modelling and activity.

National Flood Forecasting System (NFFS)
The national forecasting system currently being developed for use within the Environment Agency.

Peak (in context)
The moment during a potential flood event at which a key flood risk variable, for example river level or overtopping rate, reaches its maximum value.

Performance indicators (for flood forecasting)
Indicators may be subjective, based on opinion without an associated numerical measure, for example public perception, or a general review of numbers of appropriate warnings and false alarms. The term is sometimes also used to refer to numerical measures generated automatically alongside forecast values of flood risk variables.

Performance measures (for flood forecasting)
Measures are objective, based on a numerical evaluation not affected by forecasters' opinions, for example the mean and standard deviation of the differences between forecast and measured peak water level. The term implies a more significant evaluation than an indicator, perhaps averaged over a period of time, or focused on more severe conditions.

Physical zone (in flood forecasting context)
Atmosphere, ocean, river, nearshore, defences, inundation area and impact area.

Reliability (of forecasts)
Referring to the probability that the forecast service will function properly during potential flood events, implying availability of necessary measured or forecast input data, availability of forecasters and equipment, and a modelling solution whose calibration remains adequate during extreme conditions.

Response (to the threat)
Referring to actions that might be taken to mitigate potential injuries and damage during flood events.

Sources (in flood risk context)
Refers to the environmental loadings that might combine to produce a flood event, i.e. waves, sea level, surge, rainfall, river flow and perhaps wind.

Standby (in flood forecasting and warning context)
An internal Environment Agency alert status triggered by exceedence of one or more initial flood variable thresholds.

Storm Tide Forecasting Service (STFS)
An operational forecasting service run by the Met Office to predict surge (and tide) around the UK, targeted alerts being issued for high predicted surge and/or total sea level.

Thresholds (of flood risk variables)
Pre-computed values, exceedence of which would correspond to changes in flood behaviour (e.g. onset of overtopping or flooding to a certain depth).

Timeliness (of forecasts)
Timeliness refers to the time necessary to take effective actions in mitigation of flooding, in terms of warners, emergency services and the public being able to take appropriate actions before the onset of flooding. At present, a flood forecast is accepted as timely if associated flood warnings reach those who will have to act upon them at least two hours in which to take the necessary actions.

Triggers (for actions and flood warnings)
Triggers, either in the form of thresholds expected to be exceeded or modelling conditions being met, are signals for actions to be taken, in terms of heightened monitoring, increased alert status (e.g. standby), issuing of warnings or actions in mitigation.

Variables
Weather and ocean variables provided to flood forecasters, e.g. rainfall and offshore waves, are referred to as environmental variables. Variables predicted by flood forecasters, e.g. river level or overtopping rate, are referred to as flood risk variables.

Warner (in context)
A real English word, adopted here to refer to a person involved in preparing and issuing flood warnings, usually a member of the Environment Agency, Met Office or Emergency Services.

Warning (in context)
Referring to flood warning(s) being issued to organisation(s) outside the Environment Agency's flood forecasting and warning teams.

WRIP (flood forecasting model)
Flood forecasting system for transfer function rainfall-runoff modelling.

CONTENTS

Tables

Figures

Boxes

Appendices

PART ONE: REVIEW

1. INTRODUCTION

1.1 The purpose of flood forecasting

The purpose of flood forecasting and warning is to organise effective action to reduce potential damage and injury. To achieve this, it has to:

- *provide timely warnings*, delivered to the emergency services and to the public in time for them to take effective action in response
- *be accurate* enough in predicting flooding to encourage public action in response to warnings and to provide greater value than weather forecasts alone
- *be reliable*, in the sense of continuing to operate even in severe weather or flooding conditions
- *provide good value*, in terms of damage reduction compared to the cost of the service.

1.2 The overall context of flood forecasting, warning and response

Effective flood forecasting and warning involves:

- **Monitoring** of environmental variables that may lead to flooding, for example river flow, rainfall, tide, surge and waves, coupled with **Detection** of high values amongst the monitored variables
- **Forecasting** of possible flood incidents when one or more threshold levels are exceeded amongst the environmental variables
- **Warning**, in the sense of decision making on impending flood incidents when flooding risks are realisable
- **Dissemination** of warning messages to the public, professional partners and emergency services
- **Response** to the threat, to mitigate potential losses and injuries.

Poor performance in any one of these sub-processes would impact on the effectiveness of the overall flood forecasting and warning service. Although this report focuses on performance measures for flood forecasting, this should be seen in the context of the overall service and its ultimate objective of mitigating the potential damage caused by flooding.

1.3 Background to the project

The Environment Agency is responsible for flood forecasting and warning in England and Wales. It supports continued development and improvement of flood forecasting services, and consistent implementation across the Agency regions. To this end, development and delivery of a National Flood Forecasting System was commissioned by the Environment Agency. Prior to that work, the Environment Agency had also commissioned a series of projects to assist with the development of flood forecasting. These include:

- *forecasting extreme water levels in estuaries for flood warning* (Halcrow and Bristol University; reference Environment Agency / Defra, 2002)

- *fluvial flood forecasting for flood warning real time modelling* (Atkins; reference Environment Agency / Defra, 2003)
- *best practice for coastal flood forecasting* (HR Wallingford, Posford Haskoning and Atkins; references Defra / Environment Agency, 2003a and 2003b).

An element missing from these projects, as explicitly discussed in *best practice for coastal flood forecasting* (Defra / Environment Agency, 2003a and 2003b), is objective performance measures against which to evaluate the success or failure of forecasting services. The performance of flood forecasting within the context of the Flood Warning service is currently difficult to measure. Present approaches are at best *ad hoc* and inconsistent. Recent Agency work (Environment Agency, 2004a and 2004b) has developed a performance measurement tool to measure the benefits of the flood forecasting and warning service as a whole, but the role of flood forecasting has not been adequately defined. In addition, the apparent lack of confidence in forecast results is yet to be quantified. The need for improved and more costly modelling solutions cannot be justified if the lack of confidence cannot be quantified.

The current project was discussed during 2003, and commissioned in 2004, with a view to defining performance measures for potential adoption within the Environment Agency. Performance measures could be applied to individual forecasting models, to the separate monitoring, forecasting, warning, dissemination and response sub-processes, and/or to the flood forecasting and warning service as a whole. There are some specifically fluvial flood risk variables such as peak water level and peak flow rate; some specifically coastal such as waves and sea level; and some which would apply to both such as time of peak level, volume of flood, breaching and area flooded. Performance could, for example, be measured in terms of the following aspects of the processes and outputs, if objective parameters could be developed.

- **Accuracy** and **reliability** of forecasts – flood forecasts can only be as accurate and reliable as the underpinning general purpose weather and offshore forecasts, coupled with any additional field data assimilated in near real-time, and so need to focus on the specific prediction and consequences of flooding, keeping false alarms to a minimum.
- **Timeliness** of warnings, in terms of the capability to tune the whole service to the needs of the population at risk from flooding, and the appropriate response of the population and of the emergency services to those warnings.
- The **benefit /cost ratio** between the value of potential damage mitigated by flood forecasting, warning and response and the cost of establishing and maintaining the service (but this is a wider policy issue rather outside the scope of this report).

The recently established Flood Risk Management Research Consortium will consider sources and propagation of uncertainties, and real-time updating, in the context of flood forecasting, but there is no explicit intention to work on measuring or summarising the performance of the models. The present project and FRMRC are therefore complementary rather than overlapping.

Another complementary project on fluvial flood forecasting was commissioned and undertaken at about the same time as the present project: *Protocols for minimum standards in modelling* (W5C-021, Part 2a, contractor JBA Consulting). The purpose was to create ‘protocols’ for the creation of minimum standards in modelling within

flood forecasting. The protocols take the form of a series of statements and checklists for use at various stages within development of modelling procedures, to provide a documentary record of compliance with these minimum standards. Environment Agency / Defra (2004) introduces about 25 protocols, including Protocol 4.3 relating to assessment of fluvial flood forecasting model performance.

1.4 Objectives of performance measurement

- to provide objective, consistent and repeatable measures of the ability of flood forecasting to meet its targets, and any changes in that ability over time
- to summarise performance and value to the public, funders and regulators
- to assist in identifying and correcting weak links in individual models, regional differences and/or the overall flood forecasting and warning service
- to gain information, understanding and lessons from past experience, to assist in improving practices both locally and nationally.

1.5 Objectives of this project

The Specification went through several iterations between first being written in January 2003 and being finalised at the project inception meeting at HR Wallingford on 19 March 2004. There were three interim reports, each followed by project board review, prior to issue of a final overall report in December 2004.

The overall objectives were to produce measures to assess the performance of current (and perceived future) Agency flood forecasts, and to examine how those measures might contribute to improving the flood warning service. The specific objectives were to:

- understand the current drivers and high level targets for real-time flood forecasting and warning through international literature review
- understand how flood forecasters and warners currently use forecasts and define successful performance, identifying the limitations of these current approaches
- define the requirements for performance measurement for fluvial and tidal flood forecasting within the context of the overall flood warning service and produce a position paper for review by the Project Board
- recommend performance measures for fluvial and tidal flood forecasting for potential adoption by the Agency and produce a draft interim report on this for review by the Project Board
- outline a plan for Agency wide implementation of the recommended generic performance measures
- produce a Technical Report in current Environment Agency R&D format at the time of production, documenting all areas of the project and addressing the specific objectives above, together with recommendations for further work.

1.6 Outline and status of this report

Chapter 2 reviews current practice in evaluation of flood forecasting services. Chapter 3 discusses issues and requirements, and introduces a classification of the various physical processes and parameters to be performance measured. Chapters 4 and 5 give the recommendations of this project, in terms of definition and operation of

performance measures, implementation within the Environment Agency, and further R&D required to facilitate performance evaluation.

The Environment Agency will refine the outline implementation plan given in this report, and will continue to develop the recommendations in this report, after completion of the project. The intention is to enable significantly improved systems diagnostics and the identification of 'weak links' in the delivery of flood forecasting and warning. The longer-term aims include:

- facilitating a common approach to carrying out post-event analysis
- performance targets customised for each population at risk and each sub-process, as well as the whole flood forecasting and warning system
- facilitating appropriate feedback mechanisms to be built in through the system
- implementation into flood forecasting software.

2. CURRENT PRACTICE IN PERFORMANCE MEASUREMENT

2.1 Drivers and targets

The Environment Agency aims to provide flood warnings in sufficient time for people to take avoiding action and with a minimal number of false alarms. Key organisational drivers have included:

- Section 166 of the Water Resources Act, 1991
- the Ministerial Directive that made the Agency (and no longer local police forces) responsible for issuing warnings to the public from 1 September 1996
- development of a National Flood Warning Strategy for England and Wales in 1997
- the second phase of the National Flood Warning Dissemination Project (1997-2002)
- the Easter Floods Action Plan of November 1998 (Environment Agency, 1998) issued following the floods of Easter 1998 (and prompted by Peter Bye's review Report)
- publication of a National Tidal Flood Forecasting Joint Action Plan (MAFF *et al*, 1998) in 1998
- development of a National Flood Warning Service Strategy for England and Wales in 1999 (Environment Agency, 2001a)
- publication of "Reducing Flood Risk - A Framework for Change" in 2001 (Environment Agency, 2001b)
- the Environment Agency internal Making it Happen implementation targets
- in parallel to these are the following economic and socio-political drivers:
  - potential flood damage is increasing, in real terms, to goods stored in domestic, retail and industrial properties in flood risk zones
  - damage may be significantly reduced with adequate warning
  - disruption to traffic and public services can be reduced in times of flood if warnings are given in time to set up diversions
  - there is a greater need to give warnings in flood risk zones where mitigation works have been carried out
  - evidence of recent events confirms that the public, industrialists, public services etc now expect substantially improved flood warning services

In 2000, the Environment Agency had a number of High Level Targets for flood warning which included (Environment Agency, 2000a):

- **Reliability**: 80% success rate in provision of flood warnings
- **Residents available**: 80% success rate in the availability of the public to respond
- **Residents able**: 95% success rate in the ability of the public to respond
- **Residents effective**: 85% success rate in the ability of the public to take effective action.

The Environment Agency's Customer Charter (Environment Agency, 2001c) additionally mentions a two hour minimum warning time which is often referred to as 'timeliness'. There are also targets for the 'coverage' of the flood warning service, but this is outside the scope of this project.

Environment Agency (2004a and 2004b) re-worked these targets. The benefits estimated for the Agency's Flood Warning Investment Strategy 2003/04 to 2012/13 are based on the following performance measurement targets:

- **Damage reduction**: 40% by 2009/10
- **Coverage**: 80% by 2009/10
- **Service effectiveness**: 80% by 2009/10
- **Availability**: 80% by 2009/10
- **Ability**: 85% by 2009/10
- **Effective action**: 85% by 2009/10

2.2 Currently used and proposed definitions

2.2.1 Defra (2004)

Flood and coastal defence project appraisal guidance: Volume 6: Performance evaluation (Defra guidance note FCDPAG6) relates primarily to performance evaluation of flood defence schemes. Although the detailed methodology is not applicable to flood forecasting, much of the general discussion of performance measurement issues is relevant. It defines performance evaluation (in a way that would also apply to flood forecasting) as:

> *"a formal and periodic performance review with the aim of demonstrating value for money (e.g. Best Value), providing lessons for future management of the process under consideration and disseminating experience throughout the flood and coastal defence industry"*

In a section specifically on forecasting and warning, it proposes the following performance measures:

- **Accuracy** of the system at the sub-process interfaces between detection, forecasting, warning, dissemination and response, calculated from comparison between measured and predicted rainfall, wave height, water level etc
- **Timeliness**, defined in terms of those at risk from flooding being given sufficient lead-time before flooding to be able to benefit from the warning by taking actions to mitigate potential damage or injury
- **Coverage**, measured in terms of the proportion of those people or properties at risk from flooding receiving a warning with sufficient lead-time to take action to mitigate potential damage or injury
- **Reliability** is thought of as some function of accuracy and timeliness
- **Damage Reduction Factor** representing the proportion of the total damage that would have been caused by flooding that either was or could have been mitigated by actions following timely warnings.

2.2.2 Defra / Environment Agency (2003b)

Guide to best practice in coastal flood forecasting (Defra Project FD2206, written by HR Wallingford) notes that performance appraisal for coastal flood forecasting services is usually related to the thresholds for providing warnings, and whether or not these are

appropriate and sufficient. It suggests the following aspects of performance could be considered in connection with a coastal flood forecasting service:

- **Timeliness**
  - Can the service deliver warnings in time to help the population at risk of flooding, and in time for mitigation of losses?
  - Is the balance right between timeliness, and the extent and complexity of the modelling?
- **Reliability**
  - Will source data, people and dissemination channels be available to run the system continuously, with any additional resources needed at times of potential flood events?
- **Accuracy**
  - Is the coastal flood forecasting service significantly more accurate in predicting *flooding* than a general impression of severe sea conditions which could be gained from standard weather and ocean forecasts?
- **Appropriateness and sufficiency of forecast variables**
  - Could other variables and/or threshold levels help to determine a more precise or localised flood risk mitigation strategy?
- **Public perception and take-up**
  - Would people take any action in receipt of the proposed warnings, resulting in a reduction in the potential losses due to flooding?
  - Is the coastal flood forecasting service perceived as successful?
  - Is the cost appropriate to the potential benefits?

The report also stresses the need (at least in medium and high risk areas) for flood forecasting not only to forecast the causes of flooding but also the extent and ideally the impacts of flooding, in order to optimise mitigation of damage and injury. If this does become standard, then performance measures would need to be similarly extended to cover:

- **Sources**
  - Sources are the variables which provide the root cause of flooding, namely rainfall, tide, surge, waves (and perhaps wind and current) taken from real-time measurements and weather forecasting models.
  - Estimation of sources involves transformation of those variables from the point of measurement or forecasting through to the location(s) at which flooding might occur, using look-up tables, nearshore wave and tide models, rainfall-runoff and river flow models etc.
- **Pathways**
  - Essentially, pathways deliver the source variables to the people or properties vulnerable to flooding, so this would include interaction of the sources with the flood defences and then propagation of the flood water over the land.
  - This would involve prediction of overtopping and breaching, and then mapping of flood extent, depth and duration.
- **Receptors**
  - Receptors are the people or assets potentially vulnerable to flooding.

  - This would involve prediction of the individual people, assets or areas that would be affected by flooding, for example areas to be closed to pedestrians and/or traffic, and individual buildings liable to be flooded.
- **Consequences**
  - This refers to the 'value' of damage, injury, and social and environmental losses caused by flooding.
  - This could be applied to individual people, assets or areas, and could be re-run for a number of different emergency responses, assisting in monitoring and demonstrating the value of flood forecasting.

2.2.3 Environment Agency / Defra (2003)

Flood forecasting: Real time modelling (Environment Agency Project W5C-013/5, written by WS Atkins) reproduces two definitions of **Timeliness** used within the Environment Agency in 2000 and 2001, respectively, and offers a third definition of its own.

> *"Prior warning will be provided (two hours in general) to people living in designated flood risk areas where a flood forecasting facility exists and where lead times enable us to do so."*
>
> *"We will aim to do so at least two hours before flooding happens in areas where a service can be provided."*

In other words, the public should receive flood warnings two hours before the event, where it is both technically feasible and economically justified to offer such a service. At the same time it was understood that emergency services and Environment Agency staff would expect a six-hour warning lead-time.

> *"Timeliness expresses the expected requirements of the population at risk of flooding in terms of the time needed for effective mitigatory actions."*

The report also points out that minimum warning time is only one aspect of the total time involved in the overall fluvial flood forecasting and warning process. It recommends that as many as are relevant of the following be summed to estimate the total time involved:

- time taken for the telemetry system to poll all outstations in the catchment
- time taken to process and quality control incoming data
- time interval at which Met Office rainfall measurements and forecasts are received
- time taken for a forecasting model to run and the time interval between each run
- lead-time provided by the forecasting model(s)
- time taken for the forecaster to pass a clear forecast to the warner
- appropriateness of any trigger levels or alarms which are set including contingencies
- time taken to run additional 'what if' scenarios and interpret the results
- time taken for flood warning staff to interpret forecasts and decide whether to issue a warning
- time taken for warnings to be issued via AVM, flood wardens etc to all properties at risk.

The report also notes that whilst the Agency's definitions of **Timeliness** and **Reliability** are reasonably clear-cut, those for **Accuracy** are less so. It uses the following provisional definition:

> *"Accuracy" expresses the expected technical performance of a flood forecasting and warning system expressed in terms of appropriate criteria at interfaces (e.g. peak level reached, depth and extent of flooding etc at the interface between flood forecasting and flood warning).*

It reproduces the fluvial flood forecast **Accuracy** requirements in Table 1 from Environment Agency (2000b).

Table 1 Fluvial flood forecast accuracy requirements (reproduced from Environment Agency, 2000b)

Service level	Public	Emergency services	Agency staff
Accuracy of flood depth forecast	±0.5m	±1.0m	±2.0m
Accuracy of flood duration estimate	±3hrs	±3hrs	±3hrs
Accuracy of targeting	80%	100%	N/A
Reliability	75%	50%	50%

It notes that the level of service required may also be guided by the nature of the information required by the public and emergency services (and Agency staff) and the likely precision demanded by these 'customers'. Table 2 below shows some typical questions asked during a fluvial flood event and outlines how these might translate into accuracy requirements.

Table 2 Typical points to be addressed by fluvial flood forecasting outputs, and associated accuracy requirements (reproduced from Environment Agency / Defra, 2003)

Question	Typical requirements
When will the flooding begin?	Time at which a threshold level is reached
What depth will be reached?	The peak level reached and/or the volume of water spilling onto the floodplain
How long will the flooding last?	Times of crossing a threshold (rising and falling limb)
When can the 'all clear' be issued?	Time of dropping below a threshold
Which properties will be flooded?	Volume of flood over a threshold and location of any overtopping
Will this road/railway be flooded?	Location of flooding along the reach and timing/depths/velocities
Should temporary gates be raised/lowered?	Usually based on one or more predicted trigger levels
Should flow control structures be operated?	Time of onset of flooding (maybe several hours warning)

2.2.4 Khatibi *et al* (2003 and 2004)

These two journal papers have the titles *Research issues on warning lead-time and synergy in flood mitigation measures* and *Definition of best practice in flood forecasting*. They note that effective real-time flood management requires accurate, timely and reliable flood forecasts, and that accuracy, timeliness and reliability should therefore provide the basis for assessing the performance of flood forecasting. **Timeliness** represents the requirement in terms of a sufficient time for the forecasts to lead to an effective response to the flood risk. **Accuracy** measures the forecast performance against measured or observed conditions, in terms of peak flood level, overtopping rate, numbers of properties flooded, detection of flood events etc. **Reliability** relates to the robustness of forecasting services and their continuity even in extreme circumstances when other systems (e.g. power, communications) may begin to fail. Khatibi *et al* (2004) quote a definition of reliability used in systems science from Modarres (1993) as:

> *"the ability of an item (product, system etc) to operate under designated conditions for a designated period of time or number of cycles."*

Measures of this type assist in identifying poorly performing models or other forecasting system components. They are also useful in summarising performance to funders or to the public. Khatibi *et al* (2003) draw a distinction between lead-time and timeliness:

> *"lead-time is related to the technical dimension and it is an expression of the information extracted from real-time data in advance of the impending incident. As lead-time reduces, the information content of the impending incident increases and the uncertainty reduces but the significance of forecasting diminishes. However, the timeliness will be defined in terms of the requirements of the population at risk of flooding"*

2.2.5 Delft Hydraulics and Tessella (2004)

National flood forecasting system: Performance indicator module design (report to the Environment Agency, written by Delft Hydraulics) describes the performance indicator module of the Environment Agency's developing National Flood Forecasting System. It proposes several standard statistical measures to be applied routinely to individual forecasting modules, both during initial calibration and operationally. It proposes three types of accuracy measure, to be developed, to be applied to the overall forecasting system once it becomes operational. It discusses the need to consider lead-time, prediction of exceedence of thresholds, and prediction of time of such exceedences, in addition to routine monitoring of flood risk variables.

The present report is not constrained by measures proposed to be used within NFFS, but has the option to recommend some or all of them.

2.3 Flood warning investment strategy performance measures

Environment Agency (2004a and 2004b) are complementary internal reports describing developing methods for assessing the performance and value of the flood warning

service as a whole. The purpose is to provide the methodology to measure the performance targets in the Agency's overall strategy for the flood warning service in England and Wales. The first two measures will be adopted as annual Key Performance Indicators, and the last four measures as Business Delivery Units, namely:

- coverage
- effective action
- damage reduction
- service effectiveness
- availability
- ability.

Environment Agency (2004a) describes in general terms how these measures will be evaluated. It also gives estimates of their values for 2003/04, as percentages of their maximum possible values, and targets for improvements over the following nine years. Environment Agency (2004b) gives specific detail on how the values are estimated on a regional basis.

The success of the overall flood forecasting and warning service relies on successful Monitoring, Forecasting, Warning, Dissemination and Response. Poor performance in one or more of these elements would propagate and accumulate through to 'Response' and to the overall purpose of mitigating the impacts of flooding, resulting in a poor overall service. It is therefore essential that all the service elements aim for good performance, and that weak links be identified for improvement. This report focuses on the flood forecasting element, aiming to provide measures that can be adopted immediately and used easily within the Environment Agency's regional flood forecasting groups. Environment Agency / Defra (2003) discusses the ways that uncertainties and delays propagate and accumulate through the overall service.

Beyond describing them for the purpose of context, this report does not go into detail on aspects outside the remit of the Environment Agency's flood forecasting groups. These aspects include Monitoring (weather and ocean forecasting models and measurements) that provides input to flood forecasting, and Warning (the appropriateness of the actions and warnings) based on the output of flood forecasting.

2.4 Current practice in monitoring fluvial flood forecasting performance in England and Wales

A systematic review of practices in different regions is given in Appendix 1. Some relevant examples are given in more detail here.

2.4.1 North West Region

Lukey (2004) describes how to test and present the performance of the transfer function (TF) rainfall-runoff models run on the WRIP platform in the NW Region, based on objective analysis over a number of high flow events. The method includes separation of the performance of the model itself from the performance of the source data (measured and forecast rainfall). The parameters considered are the magnitude of peak flow, the timing of peak flow and the shape of the flow curve over a period of time. Since WRIP events have not always been saved, the method, summarised in

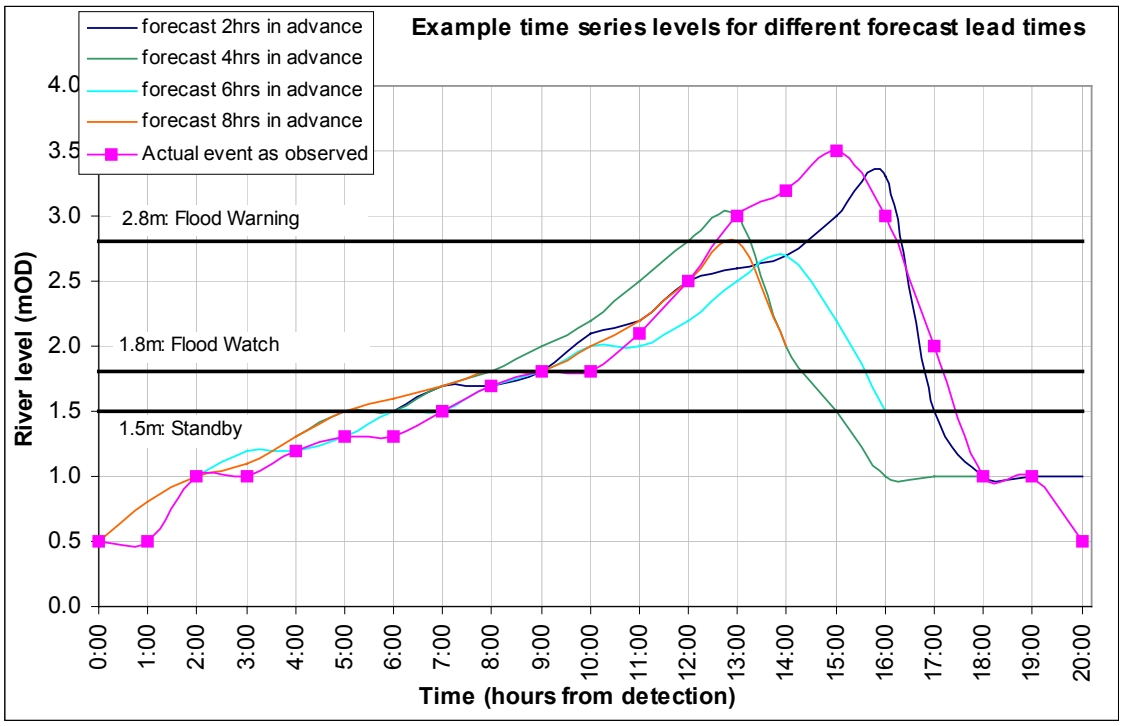
Box 1, may involve re-generation of the model performance using a simple spreadsheet TF calculation.

Box 1 Method for reviewing the performance of WRIP transfer function models in the NW Region (based on Lukey, 2004)

1. Identify events during the review period where the standby level at the site being investigated was exceeded.
2. Obtain river level and rainfall data for the event:
   - If the WRIP event was saved, obtain river level and rainfall data by loading the event and saving a model run containing the hourly data required. Save a number of model runs with various model start times, to obtain the forecast rainfall available at various times throughout the event.
   - If not saved, obtain river level and rainfall data from hydrolog and rainarc, aggregating river levels to hourly data. The rainfall data are only an approximation to those used originally. Use the WRIP rain gauge map to estimate an approximate weighting of catchment rain gauges to use.
3. Carry out a "perfect hindsight" review, using (hourly) river level and measured rainfall data as input to the TF spreadsheet along with the relevant TF parameters and flow-stage rating curve. Choose a suitable model start time (before the rainfall event commences) and compare the model predictions to measured data. Where two models exist for one location (e.g. a saturated and a dry model), plot both to see if the measured data fall within the envelope of the two models.
4. Carry out a "blind forecasting" review. Select the time at which the model might have been run (normally when the first rainfall alarm in the catchment area sounds, or when a standby river level in the catchment is first exceeded). The measured rainfall (and river levels) later than this start time are not used in the model. Plot several graphs in this way for different start times, illustrating the accuracy of the model at several different lead-times in advance of the peak.
5. Carry out an "including rainfall forecasts" review. This is a modification of Step 4, to include a combination of measured and forecast rainfall, by addition of 6 hours of forecast rainfall. Where the WRIP event was saved, use the original forecast data. Otherwise, take the first 2 hours directly from rainfall measurements (as an approximation to what the forecast would have been); the next 4 hours of measurements are either ignored, factored up or factored down to represent the expected accuracy of the forecasts.
6. Prepare a report, explaining the method followed and the data used. Include graphs comparing model performance with measured river levels, tables of the accuracy of the predicted peaks for the three types of review, and comparison with any formal forecasts issued by MFDOs. Make recommendations about model usage and re-calibration, as necessary.

Figure 1 illustrates the type of graph required in Step 6 of Box 1. The upper diagram shows progressively updated forecast water levels, for different forecast lead-times, against measured levels. The three horizontal lines correspond to water levels triggering standby, flood watch and flood warning conditions. The lower diagram shows the same measured time series levels. The five diamonds compare the measured peak level and timings of the peak, to those predicted for the different forecast lead-times (2, 4, 6 and 8 hours). The 2 hour lead-time forecast is closest to actual for

peak level and equal closest for timing of the peak level, but there is no sign of the progressive improvement that might have been expected from 8 hours to 6 hours to 4 hours.



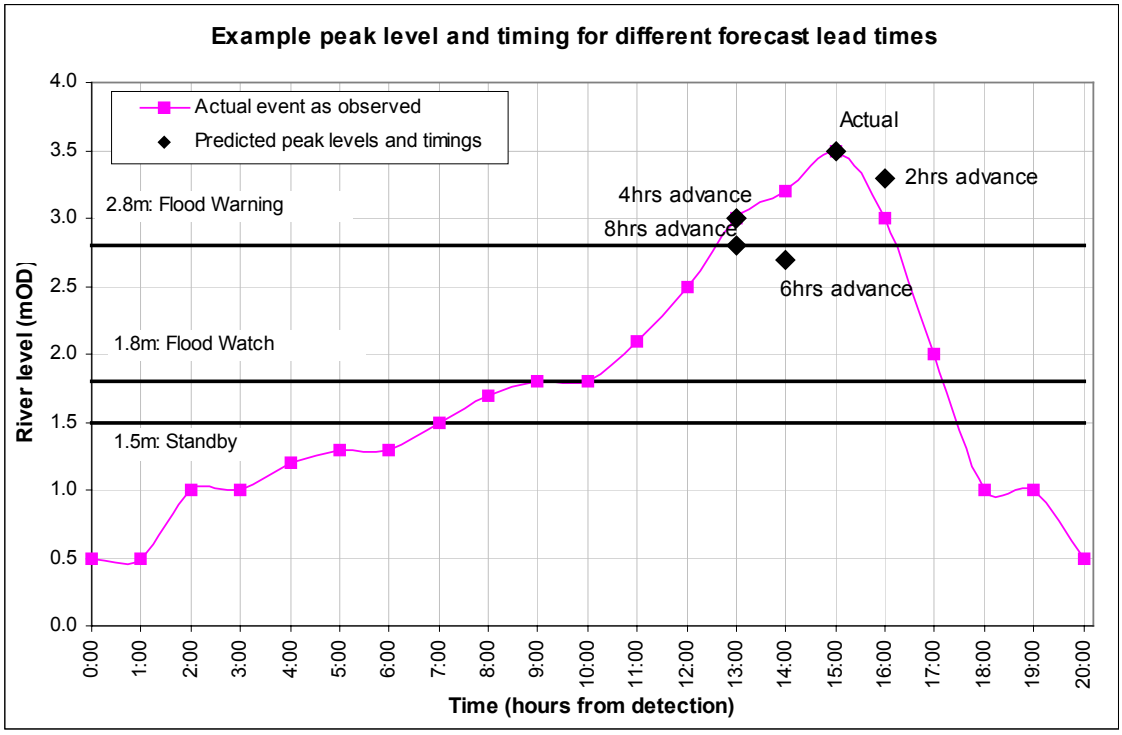


Figure 1 Comparison of model performance with measured river levels (example of format required in Step 6 of Box 1, based on Lukey, 2004)

Lukey (2003) proposes a summary performance measure in the form of a 'contingency table' described in the upper part of Table 3. The figures in the table are based on a binary (success or failure) analysis of all forecast and/or actual alert levels reached. Lukey (2003) also proposes a series of skill scores based on this analysis, as described in the lower part of Table 3. This approach has the advantage of simplicity and the potential for applicability to performance summarised over a number of events and/or a number of locations.

Table 3 Example contingency table (reproduced from Lukey, 2003)

Threshold forecast	**Flooding observed**	
	Yes	No
Yes	a	b
No	c	d

Condition	**Description**	**Definition**	**Minimum standard**
1	False Alarm Rate	$\text{FAR} = \frac{b}{a+b}$	50% or less
2	Probability of Detection	$\text{POD} = \frac{a}{a+c}$	50% or greater
Not proposed	Hit Rate (proportion correct)	$\text{H} = \frac{a+b}{a+b+c+d}$	None
Not proposed	Critical Success Index	$\text{CSI} = \frac{a}{a+b+c}$	None

2.4.2 Midlands Region

Examples of the type of information collated at present by the Midlands Region to assess flood forecasting performance for individual events were provided by Richard Cross. Figure 2 (Cross, 2004a) illustrates the February 2004 flood event at Bewdley, near Kidderminster, on the River Severn. The lower diagram shows the magnitude of the predicted peak flood, from detection of a potential event about three days beforehand, to refinement of its predicted magnitude during the two to three days leading up to the time of peak flood. The upper diagram shows the predicted timing of the forecast peak, relative to the actual peak: accurate two days beforehand but straying to 15 hours late about one day beforehand.

Table 4 (Cross, 2004b) presents a summary of the forecasting performance during the same February 2004 event, based on forecasts at about twenty locations on the River Severn made at several different times during the event. The format is comparable to that proposed by Lukey (2003), of a contingency table coupled with skill scores. The upper part of the table shows information for each of four sub-areas (US Wales, US England, LS and Severn Basin) and for each of three threshold levels (FW, FWU and SFW). The numbers in the upper left part of the table compare forecast and measured occurrences of the threshold levels (numbers in shaded boxes indicating agreement between the two). The numbers in the upper right part of the table indicate the lead-time available for these forecasts for the different sub-areas and thresholds. The figures, summarised further in the lower part of the table, indicate good forecasting accuracy, typically with greater than six-hour lead-time.

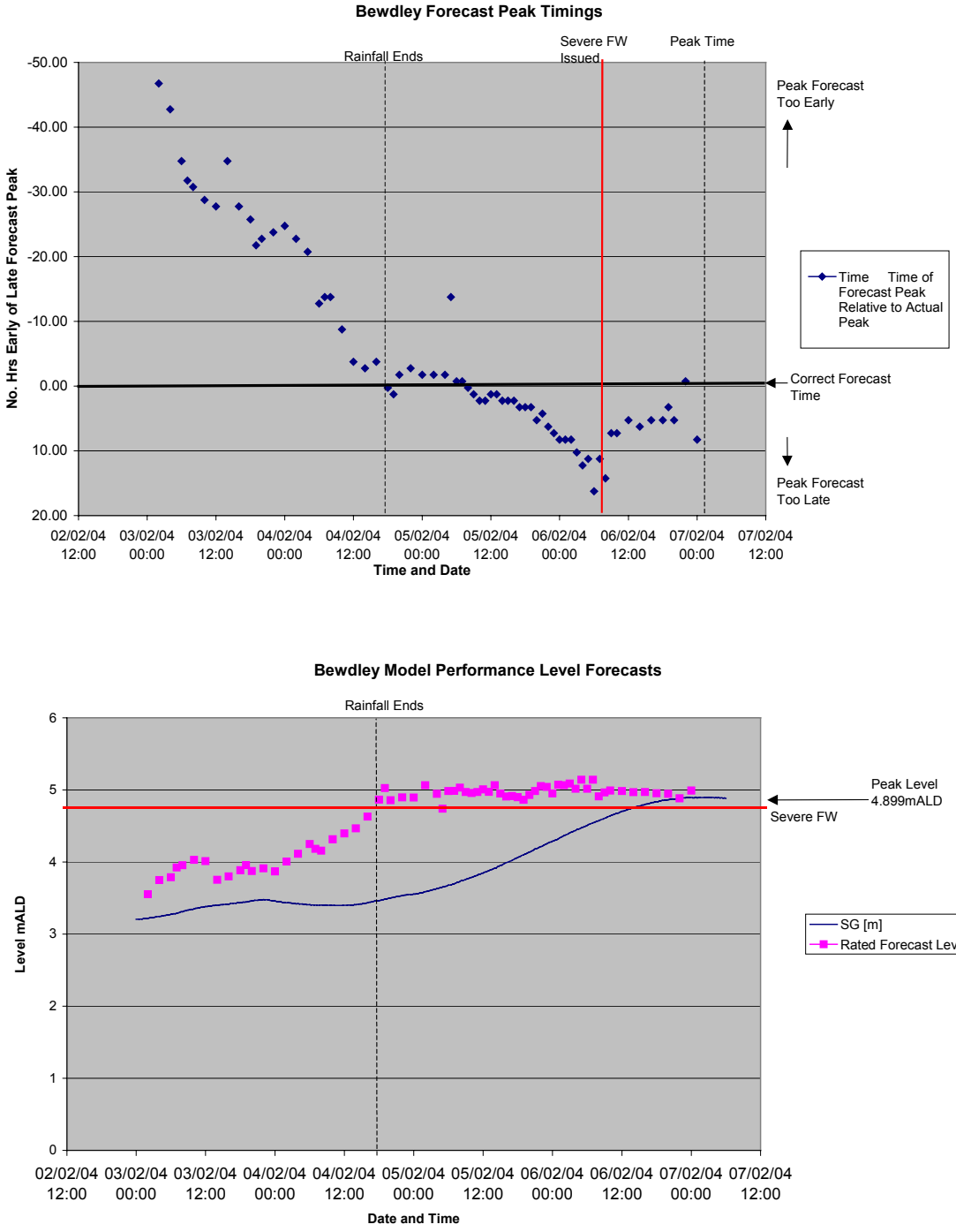


Figure 2 Flood forecasting performance during the February 2004 flood event at Bewdley, near Kidderminster, on the River Severn (reproduced from Cross, 2004a)

Table 4 Summary of flood forecasting performance during the February 2004 flood event on the River Severn (reproduced from Cross, 2004b)

	Accuracy = Total number of successes / total number of events						Timeliness - Threshold Leadtimes (Hours)					
		Actual Level			Total	Overall						Modal
Area	Warning Issued	FW	FWU	SFW	Warnings	Accuracy	<0	0 - 2	"2 - 4"	"4 - 6"	>6	Lead time
US Wales	FW	14	0	0	17	82.4%	0	4	6	2	2	2--4
	FWU	0	3	0	3	100%	1	0	0	0	1	MM
	SFW	0	0	1	1	100%	0	0	0	0	1	>6
US England	FW	11	0	0	11	100%	0	0	1	1	9	>6
	FWU	0	7	0	7	100%	0	1	3	0	3	MM
	SFW	0	1	1	2	50%	0	0	0	0	1	>6
LS	FW	3	0	0	3	100%	1	1	2	2	3	>6
	FWU	0	3	0	3	100%	0	0	0	1	0	4--6
	SFW	0	0	0	0	---	0	0	0	0	0	MM
Severn Basin	FW	28	0	0	31	90.3%	1	5	9	5	14	>6
	FWU	0	13	0	13	100.0%	1	1	3	1	4	>6
	SFW	0	1	2	3	66.7%	0	0	0	0	2	>6

MM: Multi Modal

Wales	
No. records with 1 hr forecast target	18
No. records where 1 hr target met.	13
No. records where 1 hr target missed	2
DNC's=	3
Wales timeliness =	72%
Wales Accuracy =	83.33%
England	
No. records with 2 hr forecast target	32
No. records where 2 hr target met.	28
No. records where 2 hr target missed	3
DNC's=	1
England timeliness =	88%
England Accuracy	96.88%
Total Timeliness =	**82.00%**
Total Accuracy =	**92.00%**

2.4.3 North East Region

Environment Agency (2001d) describes the assessment of performance of models used for flood forecasting in the North East Region. It details the procedures used and how the results are used in the subsequent prioritisation of forecast improvements. It is based on potential flood events, principally the Autumn 2000 event, at fifty key forecast locations.

Initial assessment is based upon the ability of the forecast to predict the time and level of the flood peak. The accurate prediction of time of threshold exceedence is also vital, as flood warnings and operations instructions are triggered at pre-specified thresholds in most cases. A method of tabulation of the performance of real-time forecast model runs has been developed. This tabulation allows the accuracy of threshold level prediction to be reviewed in addition to assessment of peak level prediction. In addition a graph is made of time against level error. Box 2 lists the variables and parameters involved in the tabulation.

Box 2 Fluvial flood variables and parameters used in comparing measured and forecast values in North East Region

Data from flood events were tabulated as follows:

1. Observed peak level
2. Observed peak time
3. Threshold level(s)
4. Observed time threshold(s) exceeded

Data from flood event real time forecasting model runs were input as follows:

1. Run number
2. Time origin
3. Overall lead-time
4. Peak value forecast
5. Peak time forecast
6. Threshold exceedence time(s) forecast

And the following data were calculated:

1. Forecast end time
2. Lead-time to observed peak
3. Peak time forecast error
4. Peak level forecast error
5. Lead-time to observed threshold(s) exceedence
6. Threshold exceedence time forecast error(s)
7. Threshold exceedence level forecast error(s)

Peak Level Forecast					
	good	6	7	8	9
	medium	4	5	6	7
	poor	2	3	4	5
			1	2	3
			poor	medium	good
			Peak Time Forecast		

Once forecast performance has been assessed, individual forecast points are scored using the matrix shown in Box 2. Scores of 2, 4 and 6 are given for poor, medium and good forecasting of peak level, and scores of 1, 2 and 3 for poor, medium and good forecasting of time of peak level, to allow ranking of overall forecast performance. The results are used to help prioritise forecast improvement works.

2.4.4 South West Region

Behan (2004) gives examples of the three types of performance measures currently applied to the WRIP models and level-to-level correlations used in the SW Region.

Tables of ‘actual’ against forecast flood watch and flood warning levels are kept for high river levels over a period of years at different locations. Results are classified as poor if less than 50% of ‘actual’ warning levels are correctly predicted, OK if 50-75%, good if above 76% and unknown if less than five events in the table. Graphs of actual against predicted peak level are kept, again over a period of years. In both cases the focus is on events rather than on continuous performance measurement. In both cases, separate dry, wet and saturated rainfall-runoff models are assessed on the same summary sheets for ease of comparison.

Level-to-level correlations are assessed over a period of years for sample locations for which water level measurements also exist. An R-squared value of 0.7 is the minimum for an acceptable correlation.

2.4.5 Thames Region

Green (2004) explains the calibration procedures used in the Thames Region for operational rainfall-runoff models. At present, there is no post-event verification, although procedures are currently being developed. Instead the models are tested at a large number of locations against a selection of calibration events. Performance is assessed with reference to the peak flow and peak water level during these events, and the lead-time for which reasonable predictions are available. The Thames (Sutton Courtenay to Hurley) HMFF model (the most developed HD model) was calibrated on four events, from which the modellers use the graphs that compare modelled with actual flows to decide how well the model performs.

In addition to modelling, level-to-level correlation is used. If the R-squared value of the correlation is less than 0.7, the correlation is not used in forecasting calculations. Travel time between the two sites is also relevant. Level-to-level correlation is of less use if the travel time between the two sites is less than 3 hours, because of the timing involved with issuing Flood Watches / Warnings). However, level-to-level information is not disregarded for travel times less than 3 hours because, as this can still be useful in making forecasting decisions. It is also noted how many points are used in the correlation because the more data points that are used, the higher the confidence in the correlation.

2.4.6 Southern Region

Southern Region does not apply performance measures in a formal way. They compile post-event reports after significant events, drawing general conclusions about the performance of the forecasting.

An 'Activity Report' is produced approximately quarterly. This compares warnings issued (and their timing) with thresholds crossed, peak levels reached, catchment conditions and time to issue on the AVM. It also looks at threshold crossings and peak levels when warnings were not issued. It gives a measure of how consistently the warnings that are issued follow the procedures and thresholds.

2.5 Current practice in monitoring coastal flood forecasting performance in England and Wales

Objective performance measurement for coastal flood forecasting services is more difficult than for fluvial flood forecasting services, for a number of reasons. The Environment Agency has less experience of coastal forecasting than of fluvial forecasting. The situation at sea is generally more confused, with high wind, overtopping spray, greater spatial variability of flooding and possibly breaching. Nearshore wave gauges are sparse, tide gauges record only part of the sea condition, and overtopping rate is rarely measured or even estimated. Also, appropriate threshold levels for flood forecasting action, in terms of overtopping rate or of some combination of waves and sea level, are less obvious, and predictions more uncertain, than for water level within a river.

All Environment Agency regions document and analyse flood forecasting. The North West and Southern Regions and EA Wales all carry out a general post-event review. North East Region forecasting is documented following duty procedures, which are used for post-event analysis. In the Thames Region, every barrier closure, warning and watch is documented. South West Region uses Defra High Level Target 2d, which is the specified national target. Thames, South West, Southern and Anglian Regions and EA Wales have coastal flood forecasting performance appraisal systems.

Methods for measuring the performance of flood forecasts, like the methods for creating flood forecasts, vary across Regions. In most instances performance appraisal is related to the thresholds for providing warnings and whether or not these are sufficient. To this end post-event appraisal is one of the most useful methods of adjusting threshold levels at which actions are triggered. Most Regions operate a system for archiving the records associated with the issuing of individual warnings, and a systematic method of recording actual flooding severity against predicted, but less consistency in recording instances when warnings were issued and no flooding occurred.

North East Region updates look-up tables every five years or whenever major deficiencies are identified. Midlands Region carries out post-event analysis after extreme events. Southern and Anglian Regions and EA Wales carry out performance appraisals after every event, to identify deficiencies in the flood forecasting system and to update warning procedures and to seek improvements in the forecast data. Thames Region carries out an appraisal after every barrier closure and flood warning, to identify whether closure was necessary. South West Region appraisals involve comparison of predicted and recorded water levels to improve calibration factors for use in further forecast water levels.

All Environment Agency Regions re-assess their trigger levels on a regular basis. The North East Region and EA Wales provided a quantitative indication of the accuracy of their coastal flood forecasting services, in terms of providing timely and effective warnings: North East Region reported 70-80%, and Wales 70%.

The TRITON system, operated by NW Region, is the best established coastal flood forecasting system within the Environment Agency. It involves offshore wave and sea level forecasts, look-up table conversions to nearshore locations and then overtopping calculations for a number of potentially vulnerable locations. Standby, flood watch and/or flood warning conditions are initiated when predicted nearshore sea levels or overtopping rates (sometimes coupled with wind speeds) exceed pre-set site-specific threshold levels. Experience to date suggests that both these threshold levels and the

transfer functions from offshore to nearshore were initially set too conservatively, resulting in too many standby and flood watch conditions.

Following some initial corrections, development of performance measures for use from 2004 onwards, and any calibration required as a result, is likely to be based around the separate sub-processes and parameters listed in Box 3.

Box 3 Coastal flood forecasting sub-processes amenable to performance assessment (based on NW TRITON but summarised here in more generic terms)

Surge forecasts
Compare peak surge forecasts with actual levels from NTS and from POL gauges. Errors from the surge model will affect everything else, and so they must be isolated and quantified.

Wind forecasts
Evaluate wind forecasts against measurements available on NTS. Any large error would help explain any discrepancies in surge and/or wave forecasts.

Offshore wave forecasts
Wave measurement devices operated by the Agency or Met Office are sparse, but where available, compare measurements offshore wave forecasts from STFS, as any errors would propagate through to inshore wave and overtopping forecasts.

Inshore wave forecasts
Site-specific inshore wave forecasts are calculated separately for each monitoring site included in the forecasting and warning service. Potentially these could be evaluated in terms of peak wave height, corresponding wave period and time of peak, for different forecast lead-times, but at present there are few inshore wave measurements to verify against.

Local sea level forecasts
Site-specific local sea level forecasts are calculated separately for each monitoring site included in the forecasting and warning service. These could be evaluated in terms of peak level and time of peak, for different forecast lead-times, by comparison with tide gauge measurements and observations of overtopping and damage.

Local overtopping forecasts
Site-specific local overtopping rate forecasts are calculated separately for each monitoring site included in the forecasting and warning service. In TRITON, maximum and mean overtopping rates are predicted, and the volume of overtopping over one high tide, and these are compared with trigger levels. (It also uses separate sea level triggers.) These could be evaluated for different forecast lead-times, by comparison with observations of overtopping and damage. (Field measurement of overtopping would be possible, but impractical; estimates of overtopping rate can be made by direct observations during an event, or subsequently from video records, at locations corresponding to those used in the forecasting model.)

Trigger levels for standby, flood watch and flood warning
The most obvious measure of flood forecasting performance is whether it correctly identifies standby, flood watch and flood warning conditions. It must neither miss events that an experienced forecaster would have recognised, nor produce too many

false alarms. As both overtopping rate prediction and appropriate trigger levels are uncertain, the trigger levels will need to be raised or lowered as experience and observational information is gathered. As part of the TRITON evaluation, observations and photographs are made whenever a flood watch condition is triggered in a forecasting model: an estimate of overtopping rate is made and an assessment of the appropriateness of the flood watch condition. This information is useful not only for site-specific refinement of forecasting models, but also to accumulate experience of appropriate flood watch levels which could be used at other coastal sites.

Some areas of England and Wales are better served than others with nearshore wave and sea level measurements, for example the Channel Coastal Observatory area between Bournemouth and Folkestone. It would seem sensible to concentrate initial validation and performance measurement trials for coastal flood forecasting services in these areas, whilst experience is gained which can then be applied in other areas.

2.6 Current practice within other organisations

2.6.1 General comments

An internet and library search was made using keywords 'performance', 'measurement', 'flood' and 'forecasting', and then using the same words coupled with particular countries in northern Europe and North America. Although many references were found, for example in USA and Australia, none suggested that other countries have made more progress than the UK in terms of objective measurement of flood forecasting performance. Most references found were in the form of commentaries and/or overlaid time series plots, giving general impressions of the level of accuracy, but without objective measures. There were several references acknowledging that performance may not be optimal, and others providing subjective assessments of the relative forecast accuracy for different lead-times. Nothing was found that seemed better than approaches already used in the UK or already in development within NFFS.

Several meteorological institutes throughout the world carry out oceanic scale operational wave forecast modelling, routinely generating data for comparison against wave measurement buoys. The scores from these analyses are frequently published (e.g. Lalbeharry, 2001) providing not only a subjective indicator of model performance but also model intercomparison.

WMO (1998) recommends various performance indicators to be applied routinely to offshore wind and wave predictions, for several different lead-times. These include monthly time series plots of forecast against measured wind speed and wave height, and corresponding root mean square error and bias statistics. As in flood forecasting, WMO (1998) also recommends a particular review of prediction of exceedence of key thresholds of wind speed and wave height.

One problem with the more conventional measures of model performance, e.g. mean, bias, scatter index (lists of statistical parameters are given, for example, in ONR, 2002 and Lalbeharry, 2001) is that small errors in the predicted times of peaks can lead to apparent large absolute errors. These types of performance indicators are therefore perhaps more use when comparing relative differences between predictions and/or different models.

To assess the errors in timing it may be useful to carry out cross-correlation analyses between the predicted and measured data for different lags, to provide an indication of whether the predictions are systematically early and/or late.

2.6.2 UK Met Office

For its own calibration and validation purposes, the Met Office periodically compares measured and predicted winds, waves and surges at locations where measured data are available. The comparisons are in terms of time series plots, differences between mean values over a period of time, and standard deviations of differences between predicted and measured values. These latter summary figures do not discriminate between high values and less important records, and need to be interpreted with some care. Also, the standard deviation of differences comes from a combination both of magnitude and of timing differences, which cannot then be separated out.

Met Office (2004) was prepared for wider circulation, and contains some summary statistics for the accuracy and timeliness of the Storm Tide Warning Service sea level predictions. In addition to some descriptive text on particular events, these are presented for different parts of the UK coast in the style sometimes called 'contingency tables'. These are given for:

- **Alerts performance**: This consists of a table of occurrences of alert levels, near-misses and non-events for each of four classes (defined by lead-time) of alerts issued and one additional class of no alert issued.

- **Forecasting accuracy**: This consists of a table relating to peak surge levels during actual and/or forecast occurrences of alert levels, listing mean, maximum, root mean square and standard deviation error averaged over all events during a period of time.

- **Timeliness**: This consists of a bar chart of the number of alerts issued against the number of near-misses and alert levels measured, for different alert lead-times.

Examples of each type, reproduced from Met Office (2004), are shown in Table 5, Table 6 and Figure 3. A difference from the approach proposed by Lukey (2003) is that 'near-miss' is recorded separately to 'miss', a near-miss indicating that the recorded sea level did not reach the alert level but came within 20cm of it.

It is not clear whether the concept of a near-miss would be directly helpful in the context of this project, but the near-miss category could perhaps be replaced by 'intended alert level not reached but physically significant lower alert level reached'. This is implicit in the approach used in the Midlands Region (Cross, 2004b) where different physically meaningful thresholds are used in the same contingency table.

Performance measures of this type might be suitable for adoption into coastal flood forecasting, where precise measured data are not usually available. Met Office (2004) also calculates a 'skill score', representing the percentage of cases of alert levels or near-misses being measured, when the prediction level was within 20cm of the measured value.

Table 5 Example of the 'alerts performance' contingency table used by STFS (reproduced from Met Office, 2004)

Alert issued?	Major	Alert level reached?		
		Yes	No — near miss	No
Yes ≥12 hrs	0	2	23	9
Yes ≥8 hrs and <12 hrs	0	5	44	22
Yes <8 hrs and ≥4.5 hrs	0	0	3	1
Yes <4.5 hrs	0	1	1	0
No	0	1	4	18,240

Table 6 Example of the 'forecasting accuracy' summary table used by STFS (reproduced from Met Office, 2004)

Location		Average number of events	Mean error	Average max. positive error	Average max. negative error	RMSE	SD of errors
All ports	2003/4	19	–0.07	0.24	–0.37	0.18	0.16
	mean	25	–0.04	0.27	–0.37	0.18	0.15
Divs 1 to 5	2003/4	26	–0.07	0.37	–0.34	0.18	0.17
	mean	25	–0.07	0.21	–0.38	0.16	0.15
Divs 6 and 7	2003/4	11	–0.02	0.22	–0.23	0.14	0.14
	mean	25	–0.04	0.23	–0.29	0.15	0.13
Div 8	2003/4	16	–0.14	0.30	–0.65	0.28	0.24
	mean	23	–0.09	0.34	–0.51	0.23	0.20
Divs 9 and 10	2003/4	21	–0.07	0.14	–0.31	0.15	0.12
	mean	29	0.01	0.31	–0.35	0.18	0.15

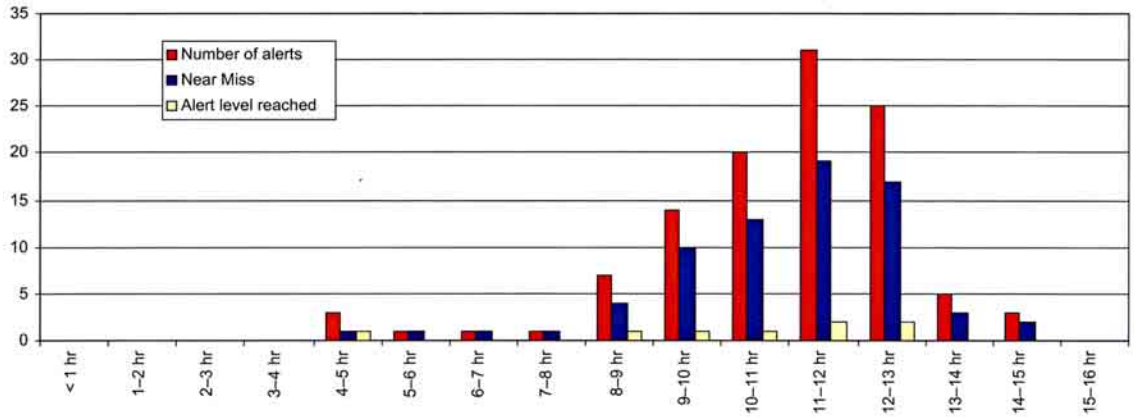


Figure 3 Example of the 'timeliness' bar chart used by STFS (reproduced from Met Office, 2004)

2.6.3 HR Wallingford Group

An example monthly performance report from a Wallingford Software FloodWorks nearshore wave forecasting system in Egypt is given in Table 7. (The equivalent Danish Hydraulics Institute and Delft Hydraulics products are MIKE Flood Watch and DELFT-FEWS.) The table shows that Mean Absolute Error, Bias, Root Mean Squared

Error and Over Prediction Ratio tend to increase with forecast lead-time. OPR is the result of a binary count of the number of over- and under-predictions. OPR = 0.632, for example, would mean that (excluding exactly correct predictions) 63.2% are over-predictions, whilst OPR = 0.5 would indicate equal numbers of over- and under-predictions.

Table 7 Example monthly accuracy statistics for forecast significant wave heights (based on Wallingford Software FloodWorks)

Measure	Average forecast lead-time (hours)										
	6	18	30	42	48	60	72	84	96	108	120
MAE (m)	0.259	0.241	0.230	0.249	0.303	0.356	0.406	0.388	0.490	0.547	0.567
BIAS (m)	0.067	0.055	0.024	0.070	0.166	0.185	0.172	0.169	0.278	0.377	0.380
RMSE (m)	0.330	0.313	0.320	0.348	0.416	0.504	0.549	0.533	0.680	0.825	0.934
OPR	0.632	0.607	0.564	0.571	0.663	0.601	0.595	0.571	0.607	0.675	0.675

Again provided measured data are available, event prediction accuracy can be used as an alternative performance indicator. Events can be defined, for example, as used in FloodWorks (2004): peak over a threshold, peak and rate, or a total over a period of time. Performance of the forecasts can then be measured in terms of events, e.g. if and when events were forecasted and the timeliness of such forecasts. This type of indicator can sometimes be a more useful measure since it is more representative of times when warnings might be issued or actions taken, and secondly it does not necessarily rely on the absolute accuracy of the forecast parameter.

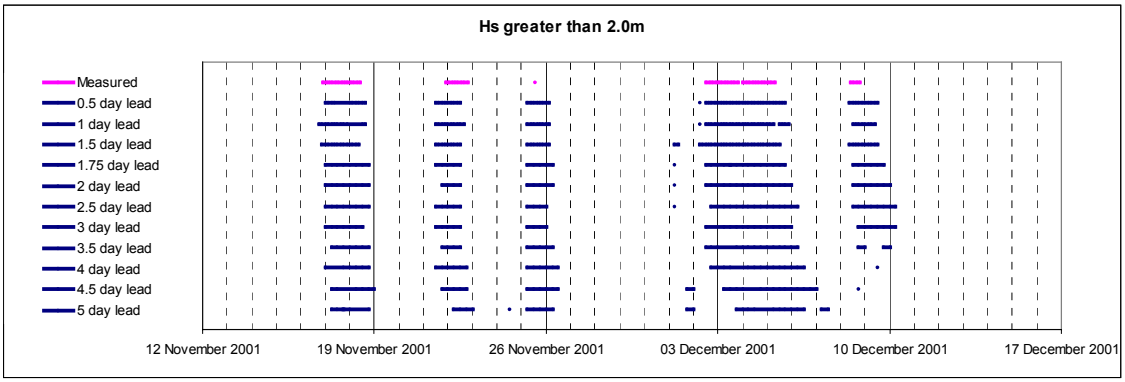


Figure 4 Example of performance indication based on nearshore significant wave height (H_s) event prediction (reproduced from Abernethy *et al*, 2004)

Figure 4 shows an example bar type graph devised to illustrate event forecast performance (Abernethy *et al*., 2004). This figure shows events (in this case significant wave heights above a predetermined threshold) shown as (blue) lines for various lead-times. These include a 5-day lead forecast in the lowest line to a 0.5-day lead forecast, with the upper (purple) line for events based on measured wave conditions. This type of presentation gives an immediate, but subjective, impression of performance. Further analysis of event prediction can also be carried out to provide objective results.

3. REQUIREMENTS FOR PERFORMANCE MEASUREMENT

3.1 The challenge of performance measurement

3.1.1 Why is performance measurement necessary?

As with any investment or ongoing commitment, it is desirable to test the results from time to time, to assess the value of the investment and to determine whether any change in strategy is needed. The need for such assessments is higher where the results of the investment are intangible or subjective (e.g. public reassurance, potential damage mitigated) or where some of those results might have been achieved without that investment (e.g. from weather forecasting). Performance measurement also provides a means by which developers and operators can identify problems and assess the relative merits of alternative approaches.

3.1.2 Who is performance measurement for?

- those directly involved in flood forecasting, e.g. developers and operators, to assist with incremental improvements to forecasting services
- the funders and regulators, to assist in assessment of meeting of targets and value for money
- the public, who use and pay for the service.

3.1.3 Who would apply performance measurement?

Flood warning in England and Wales is the responsibility of the Environment Agency. Both fluvial and coastal flood forecasting are organised within the Agency Regions, with some central guidance on best practice and procedures. As with flood forecasting, performance measurement would be applied by the forecasting groups within individual Environment Agency Regions, again with central guidance.

3.1.4 What could performance measurement help to achieve?

- identification of weak links in individual processes within the overall flood forecasting service
- improvements in modelling and forecasting processes
- demonstration of the value of forecasting and the effectiveness of improvements
- public acceptance of the benefits of flood forecasting
- continuity of funding for flood forecasting.

3.1.5 How does sub-process performance relate to overall service performance?

The overall flood forecasting and warning service is made up of multiple modelling (e.g. weather, rainfall-runoff, river, ocean, nearshore, flood defence and inundation models) and management (i.e. monitoring, forecasting, warning, dissemination and response) procedures. As most of these procedures run in sequence, any errors, uncertainties and delays will tend to accumulate as they propagate through the system. (This is not necessarily the case if calibration of the service is based only on its end-product, or if additional real-time observational data are added between service elements).

Little insight is currently available on the relationship between the performance of the flood forecasting sub-processes and the flood warning service as a whole. Although it may be interesting and feasible to make small improvements to the quality of the river numerical models used in fluvial forecasting or the nearshore wave transformation models used in marine forecasting, any improvement in the forecast end-product (mitigation of potential losses) may be small. Greater improvement may derive, for example, from increasing the resolution in the source forecasts (rainfall, surge, waves) or improving warning dissemination methods. In other words, weak links need to be identified and prioritised for replacement or repair.

3.1.6 What forecasting processes and parameters might be measured?

The types of processes and parameters are listed below, those in brackets being outside the scope of this report.

- (forecasting of rainfall or sea condition inputs to models)
- detection of potential flood events, and non-detection of non-events
- forecasting of water level, and overtopping rate and volume
- (forecasting of flood extent, depth and duration)
- (the interface between flood forecasting and warning decision making)
- (the relationship between defined flood thresholds and the onset of flooding)
- (the actions of warning recipients).

Performance at different lead-times is useful, and the specific times focused upon could correspond to target times, i.e. the normal 2-hours lead-time, 6-hours for major incidents and 8-hours for demountable flood defences. (Demountables in this context refers to planned temporary defences, placed on top of the permanent flood defences during times of potential flooding, and 'demounted' when the period of risk has passed. Examples are sandbags, gates in parapet walls on sea defences, and temporary vertical structures erected on permanent river defences.)

3.1.7 What types of performance measures are needed?

Ideally the measures should be objective and generic, and should not be site-specific, time-specific or subject to forecaster judgment. In some instances, particularly in coastal flood forecasting, these conditions may be impossible to meet because of the lack of appropriate field measurements.

Some measures need to be specific to particular sub-processes to assist in identification and replacement of weak links in models and procedures. Others need to be more general measures of the entire flood forecasting and warning service, for dissemination to non-specialists. Different levels of performance measurement might be nested inside each other, e.g. 'nearshore wave height' nested within 'overtopping', and 'overtopping' within 'coastal flood forecast threshold level'. The pattern of performance through these overlapping measures would help to track development and propagation of uncertainties through the various models and procedures.

At present, flood forecasts are deterministic. In the future, risk-based or ensemble forecasting might be introduced, involving different forecasting models and/or multiple runs of the same model using different input parameter values. This would offer the

possibility of using additional performance measures, perhaps based on the proportion of time that actual occurrences fall within the range predicted by the model(s).

The exact nature of the performance measures used may depend on the particular forecasting models used in different areas, the associated variables, parameters and thresholds used, and the measurements or observations available for comparison. For example, if overtopping rate were either not predicted or not recorded, there would be little point in its being subject to a formal performance measure. Instead, one would look to the accuracy and suitability of predicted threshold levels, based on other flood risk variables.

3.1.8 What types of performance measures are currently used?

Existing performance measures are reviewed in Chapter 2. There are several indicators related to the overall flood forecasting and warning service for England and Wales, used to assess the value of the Environment Agency's overall investment strategy in the service.

Most flood forecasting services are assessed, at least on a post-event basis, in terms of the suitability of any warnings issued (or not issued). This helps to determine whether appropriate people (emergency services, Agency staff, broadcasters, the public) were warned to take actions (e.g. flood watch, demountable defences, movement of goods, evacuation) appropriate to the event (e.g. false alarm, near-miss, small flood, severe flood). This type of assessment is related to prediction of exceedence of various pre-defined thresholds, some of which will be triggers for warnings and/or actions, and whether or not those thresholds are representative of the actual flood risk to people and property.

Some flood forecasting services are assessed against measured or observed flood risk variables. These variables include both some of the source forecasts (wind, rainfall, surge, waves) and some of the specifically flood-related variables (river flow, river level, nearshore sea conditions, overtopping rate, breaching, time of flood, flood area). Reviews following actual flood events are carried out in most Agency Regions, and in some cases also following warnings not accompanied by actual flooding. In some Regions, this type of assessment is conducted for different lead-times, to assess any change in the accuracy of the time and severity of flooding as the event approaches.

3.1.9 What are the current deficiencies in performance measurement?

Performance measures, where they are used, were developed independently within different Environment Agency Regions, and continue to be developed as experience is gained. They are therefore not consistent across Regions or through time, and would not be amenable to external audit. At present, it would be difficult to identify specific weak links, or to compare performance between Agency Regions or between fluvial and coastal forecasts. It is therefore desirable to develop specific measures and to apply them in a consistent manner.

A specific problem with existing measures is that the primary measure used is R^2 which is able only to evaluate linear relationships between variables and is insensitive to additive or proportional differences between simulations and observations. High values

of R^2 can be obtained even when model simulated values differ considerably in magnitude.

In addition to this limitation, the R^2 measure is also more sensitive to outliers than to observations near the mean, a fact which also applies to any measure (including the Nash-Sutcliffe Statistic) which involves a squaring of the difference terms.

The above issues are discussed in greater detail in Legates and McCabe (1999).

3.1.10 What are the immediate problems?

The most immediate scientific problems relate to the lack of direct measurements of flood parameters (too few instruments and too few flood events). Although rivers are reasonably well served with measurements of flow rate and level, the lack of objective measurements of sea conditions and overtopping means that routine comparison of measured and predicted coastal flood risk variables may be difficult. This would also cause difficulties for initial validation and setting of appropriate flood risk thresholds within coastal flood forecasting models.

Even where measurements of flood risk variables are available for model performance testing, without information during potential flood events, one cannot be sure that numerical models will perform to the same level in such conditions. (This refers back to reliability and whether the flood forecasting models and the overall forecasting and warning service work effectively during flood conditions.)

There may also be administrative problems associated with the need for each of the different forecasting teams to undertake the same developments in parallel in a consistent way across the different Environment Agency regions.

3.2 Immediate requirements for performance measurement

3.2.1 The purposes and nature of performance measurement

The purpose of flood forecasting is summarised in Section 1.1. The purposes of performance measurement are to:

- provide objective, consistent and repeatable measures of the ability of flood forecasting to meet its targets, and any changes in that ability over time
- summarise performance and value to the public, funders and regulators
- assist in identifying and correcting weak links in individual models, regional differences and/or the overall flood forecasting and warning service
- gain information, understanding and lessons from past experience, to assist in improving practices both locally and nationally
- improve the effectiveness of flood forecasting, with a clear focus to maintain or improve lead-time for the accurate, timely and reliable delivery of flood forecasting and warning services.

The Environment Agency's High Level Targets for flood warning are summarised in Section 2.1.

3.2.2 The scope of performance measures

The five sub-processes of a flood forecasting and warning service are listed in Section 1.2. Although this report is concerned primarily with performance measures for **flood forecasting**, it also needs to consider the interfaces with the two neighbouring processes of **monitoring / detection** and triggers for **warning**. It is also important to remember that poor performance in any one of the sub-processes would impact on the effectiveness of the overall flood forecasting and warning service.

A subtle distinction can be drawn between physically based *thresholds* related to likelihood of flooding, and *triggers* for different actions by flood forecasters and warners. An example use of a threshold would be where a flood warning level on telemetry corresponds to the level when flooding to properties is expected to begin (i.e. a flood threshold). Forecasting performance could be measured by ability to predict the exceedence of the threshold at different lead-times (as well as whether the threshold is correctly set in the first place). An example use of trigger levels would be where levels are used as alarms in telemetry (e.g. triggering a standby status). These levels are intended to initiate an action such as commencement of forecasting. In addition to ability to forecast such levels correctly, a level could be assessed in terms of its ability to give sufficient lead-time to act, without too many false alarms leading to too much unnecessary effort.

3.2.3 Classification of the physical processes and parameters

A four-level classification of the physical processes and parameters involved in forecasting and warning is introduced in Table 8, representing:

- **the general aspect of the flood forecasting and warning service**
  - **the physical zones involved within each aspect**
    - *the variables involved within each zone*
      - - the parameters to be performance measured for each variable.

Table 8 Classification of the physical processes and parameters involved in forecasting and warning

Weather forecasting	
Atmosphere	
Rainfall	
	– Peak intensity over a fixed period of time, e.g. 1 hour
	– Time of peak rainfall intensity
	– Shape of the rainfall curve over time, e.g. 20 hours
Wind	
	– Peak speed over a fixed period of time, e.g. 1 hour
	– Time of peak wind speed
Ocean	
Sea level	
	– Peak surge near the time of an event
Waves	
	– Peak wave height over a fixed period of time, e.g. 1 hour
	– Peak wave period over a fixed period of time, e.g. 1 hour
	– Time of peak wave height

Flood forecasting
River
River flow
– Peak flow over a fixed period of time, e.g. 1 hour
– Time of peak flow
– Shape of the flow curve over time, e.g. 20 hours
River level
– Peak level near the time of an event
– Time of peak level
– Occurrence of threshold crossings
– Time of threshold crossings
Nearshore
Sea level
– Peak surge near the time of an event
– Time of peak surge
– Peak sea level near the time of an event
Waves
– Peak wave height over a fixed period of time, e.g. 1 hour
– Peak wave period over a fixed period of time, e.g. 1 hour
– Time of peak wave height
Defences
Overtopping
– Peak overtopping rate over a fixed period of time, e.g. 1 hour
– Volume of overtopping during an event, e.g. over high tide
– Time of peak overtopping (fluvial only)
Breaching
– Occurrence of breaching
– Location of breaching
Inundation
Flood mapping
– Occurrence of significant flooding events
– Time of onset of significant flooding (fluvial only)
– Area extent of flooding
– Depth of flood water
– Duration of flooding
Threshold levels
River
River level
– Occurrence of different threshold levels
– Appropriateness of threshold levels to actual risk
Nearshore
Overtopping rate (and/or other, e.g. sea level and wind)
– Occurrence of different threshold levels
– Appropriateness of threshold levels to actual risk
Inundation
Flood probability
– Occurrence of standby, flood watch and warning levels
– Appropriateness of threshold levels to actual risk

For the purposes of testing **Accuracy**, any or all of these variables and parameters (and any formal forecasts issued) could be compared with equivalent measured or observed (or even inferred during the aftermath) values. This could be done for each flood forecasting location of interest and for a range of forecast lead-times from about two hours to two days. For the purposes of testing **Timeliness**, the times at which relevant variables and parameters begin to take settled forecast values could be compared with the times at which they would be required in order to deliver timely warnings. For the purposes of testing **Reliability**, the proportion of time (weighted towards more severe weather conditions) that any or all of these variables and parameters meet specified accuracy and timeliness constraints could be calculated.

For the purposes of this report, a sub-set of the processes and parameters listed in Table 8 is identified, corresponding to the responsibilities of Environment Agency flood forecasters.

Although essential to the overall success of a flood forecasting and warning service, 'Weather forecasting' and 'Appropriateness of threshold levels to actual risk' are not considered further, as performance is not the responsibility of Environment Agency flood forecasters. Similarly 'Breaching', 'Inundation' and 'Impacts' are not at present forecast (or measured) explicitly. It would not be helpful to develop performance measures for these aspects now, but merely to mention them as being of possible interest if flood forecasting develops in that direction in the future.

That leaves 'River', 'Nearshore', 'Overtopping' and 'Occurrence of different threshold levels' to be developed further in this report. The lines containing these aspects are shaded in Table 8, and are taken forward to Chapter 5, where specific recommendations for performance measurement are given.

3.3 Longer-term aspirations for performance measurement

This section raises a few issues beyond the immediate requirements and implementation plans, but which might re-visited as part of any longer-term refinement of flood forecasting and performance measurement.

3.3.1 Measurements at potential flood locations

An obvious problem in performance measurement (and in model validation) is the lack of direct measurements of flood parameters: too few flood events and too few instruments at the location(s) of flooding. Measurement at flood risk location(s) of variables directly representative of flooding would also be useful for real-time updating of forecasting models, even when flooding is not likely. (One exception to the shortage of measurements is for normal (in-bank) flow in rivers, for which there are often many years of data in most rivers.)

Additional variables which might be measured routinely (or observed during flood conditions) include nearshore wave and tide data, coastal and fluvial overtopping rates, flood extent and depth. These would all help towards objective measurement of the performance of different aspects of flood forecasting models. Even for flood risk variables not explicitly forecast at present, or for areas not served by site-specific flood

forecasting models, measurements during extreme conditions would be potentially valuable in calibrating and validating forecasting models yet to be implemented.

3.3.2 Uncertainty propagation

Understanding and quantification of the generation and propagation of uncertainties through the various sub-processes of the flood forecasting and warning service could be useful in performance measurement and identification of weak links. However, this is probably a refinement to be considered later, when normal performance measures have been implemented, and following further research into the benefits of understanding uncertainty propagation.

This might eventually be taken forward to a fully risk-based approach, in which ranges of variables would move through the forecasting processes, offering the possibility of a more complex risk-based performance measurement scheme.

3.3.3 Widening of the scope of flood forecasting

A fully developed flood forecasting service not only needs to forecast the causes of flooding but also the extent and ideally the impacts of flooding, in order to optimise mitigation of damage and injury. If this does become standard, then performance measures would need to be extended (beyond Sources) to cover:

- **Sources**

Rainfall, tide, surge, waves (and perhaps wind and current), both at the location(s) of measurement and/or weather forecasting, and after any spatial transformation carried out as part of the flood forecast modelling.

- **Pathways**

Overtopping, erosion and breaching, and then mapping of flood extent, depth and duration.

- **Receptors**

The assets potentially vulnerable to flooding, for example individual people, buildings or areas that would be affected by flooding, perhaps in the form of areas to be closed to pedestrians and/or traffic, and individual houses liable to be flooded.

- **Consequences**

This ‘value’ of damage, injury, and social and environmental losses caused by flooding, which could be re-run for a number of different emergency responses, demonstrating the value of flood forecasting and of alternative responses.

PART TWO: RECOMMENDATIONS

4. DEFINITIONS FOR PERFORMANCE MEASURES FOR FLOOD FORECASTING

4.1 Key points carried forward from PART ONE

4.1.1 What is the purpose of flood forecasting?

Flood forecasting is an essential component of an overall flood forecasting and warning service comprising monitoring, **forecasting**, warning, dissemination and response. The purpose of the service is to organise effective action to reduce potential damage and injury. To achieve this, flood forecasting needs to be timely, accurate, reliable, cost-effective and well integrated into the overall service. In terms of day-to-day delivery, the main job of flood forecasters is to predict when certain pre-determined thresholds of certain flood risk variables at certain locations will be exceeded, with sufficient time for appropriate actions to be taken in response to the threat.

4.1.2 How is flood forecasting operated?

Flood forecasting is run by the eight Environment Agency Regions. Fluvial flood forecasting is well developed in all eight regions. Except in the North West Region, coastal flood forecasting is less well developed then fluvial. Over the coming few years, all regions will adopt the National Flood Forecasting System, which will bring consistency to the procedures and information used in both fluvial and coastal flood forecasting.

4.1.3 What is the purpose of this project?

The purpose of PART ONE of this report is to understand the context of flood forecasting, and the issues, requirements, present situation and options available for objective performance measurement. The purpose of PART TWO is to recommend performance measures for flood forecasting to be developed and implemented for consistent use throughout the Environment Agency.

4.1.4 What does performance measurement deliver?

Performance measurement delivers objective and consistent information on the performance of various models and procedures. A wide array of performance measures can be used by developers and forecasters to help in identification of weak links in source data, modelling and procedures, to be prioritised for improvement. Simpler summary measures provide a comparison between different implementations, and are used to demonstrate value to funders and to the public.

4.1.5 What processes, parameters and measures are involved?

Fluvial flood forecast outputs include peak water level, peak flow rate, time of peak level and volume of flood; coastal flood forecast outputs include sea level, wave height and overtopping rate. The range of processes and parameters relevant to a flood forecasting and warning service are classified in Table 8. The sub-set of processes and parameters within the scope of this report are shown shaded in Table 8.

4.1.6 What types of performance measure are used?

Quantitative mathematical measures include maximum error, mean error, bias, standard deviation and lead-time error. These are produced continuously as part of the forecasting process, and also at intervals to look at performance over a period of time. Simpler summary statistics of events (whether forecast and or actual) are produced in the form of 'contingency tables', say quarterly or annually, for non-specialists.

4.1.7 How will the recommendations of this project be implemented?

The obvious way forward, at least where time series measured data are available for comparison, is to adopt (and possibly add to) the measures already intended to be included within the National Flood Forecasting System. Some additional summary performance measures will also be needed.

4.1.8 How will measurements and observations be used?

Local measurements of flood risk variables, such as river levels, sea levels, nearshore wave conditions, overtopping rate and extent of flooding are an essential part of flood forecasting. They are used for real-time updating of forecasting models, that is for correction of 'nowcast' conditions used as starting conditions for a 'forecast'. They are also used for periodic model calibration and in evaluation of forecasting performance.

4.1.9 A note on definitions

From here on, the definitions given in the glossary at the start of this report will be assumed to apply.

4.2 Definitions of performance measures for fluvial forecasting within NFFS

This project is not constrained to using only parameters and performance measures available within the National Flood Forecasting System, developed primarily for use in fluvial flood forecasting. However, as these will be implemented within NFFS, it is sensible to focus first on these parameters and measures, and to consider whether they are appropriate and sufficient.

4.2.1 Classification of performance measures

The classification of different performance measures for particular measured data will depend on the nature of those data rather than the physical measurement location. For example, if a measured time series (or a time series derived from measured data) is available for comparison with a simulated time series, then the same measures should be applicable whether the data relate to river levels, flood plain levels, flows or levels behind coastal defences.

Time series data is the most common form of data available to assess performance measures but in the future different types of comparative data will become available both in real-time and to inform post-event analysis. The current functionality of the NFFS can address time series comparison but not the following:

- Time series of time-stamped inundated areas from aerial survey
- Single time-stamped image of inundated area
- Single post-event maximum inundated area
- Non-equidistant series of time-stamped observed levels at multiple locations
- Single peak level at multiple locations from post-event wrack mark survey
- Single peak levels at random locations from observers

It is worth noting that performance measures are available in the scientific literature to address simulated and observed inundated areas, and also that NFFS can support the processing of data from simulated inundated area predictions. It cannot at present support the automatic processing of data from aerial survey.

4.2.2 Summary of the attributes of the NFFS performance measures

A brief summary is given below of each of the quantitative measures of time series analysis intended to be available within NFFS.

Bias

The bias error is a simple summation of the errors at a given set of N sample times, divided by the number of sample times.

The positive aspect of this measure is that it is a good measure of the relative average magnitudes of the observed and simulated time series. If the error is zero, the average magnitudes are the same.

The primary problem of this measure is that positive and negative errors cancel out so that even if the error is zero, the time series may not be coincident.

Mean Absolute Error (MAE)

Mean Absolute Error differs from bias in that the absolute values of the individual error components are used rather than the simple difference.

The advantage of this measure is that if its value is close to zero, then the observed and simulated time series are indeed coincident.

Mean Square Error (MSE)

Mean Square Error is simply the average of the square of the individual differences at each sample time. It has the same advantage as MAE, in that if the value is close to zero, then the observed and simulated time series are coincident.

It has the disadvantage that differences away from the mean values are amplified more than for the MAE.

Nash-Sutcliffe Efficiency (NSE)

The Nash-Sutcliffe Efficiency (NSE) statistic is a ‘goodness of fit’ estimate which measures the relative magnitude of the residual variance (‘noise’) to the variance of the

flows ('information'). The NSE indicates how well the plot of observed versus simulated data fits the 1:1 line. The optimal value of NSE is 1.0, and values should be larger than 0.0 to indicate 'minimally acceptable' performance. A value equal to 0.0 indicates that the mean observed flow is a better predictor than the model, but in principle the statistic can vary between -∞ and 1.

R^2 (also called the Coefficient of Determination)

R^2 is the square of Pearson's product moment correlation coefficient r (i.e. $r^2 = R^2$). It is a measure of the proportion of variability explained by, or due to the regression (linear relationship) in a sample of paired data. It ranges between 0 and 1.

The advantage of this error measure is that it is well used and understood by practitioners - both hydrologists and engineers. It is also used in several existing calibration environments such as TSCAL and WRIP. The disadvantages are described in Section 3.1.9.

In general terms a high value of NSE should correlate to a high value of R^2 but, importantly, a high value of R^2 may not correlate to a high value of NSE and does necessarily indicate that a good simulation has been achieved.

4.2.3 Comments on observed data

It is important to realise that real-time telemetered data are subject to errors, some of which may be predictable such as an ultrasonic flow gauge 'flatlining' once a particular value has been reached. Some errors may not be predictable, for example blocking of an inlet pipe, resulting in a particular measured time series not being able to assess system performance at a given gauge location.

NFFS has facilities automatically to edit observed time series and this may render it usable for performance monitoring even if certain points have been removed due to exceedence of soft or hard limits on data ranges or rate of change of data. Nevertheless, criteria to be specified should be satisfied by measured data in terms of their possible use for performance measurement.

4.2.4 Whole hydrograph prediction

The first and simplest application of the proposed NFFS performance indicator module is in module calibration. This is done by comparing two time series, over a configurable duration, one being the estimated series and the other being the reference time series.

The time series are compared using a number of performance indicators. $\hat{x}_i$ is the estimated value, x_i is the measured value, and N is the number of data points. $\overline{x}$ is the mean of the measured values.

Bias (BIAS)

$$BIAS = \frac{1}{N}\sum_{i}^{N} \hat{x}_i - x_i$$

Mean absolute error (MAE)

$$MAE = \frac{1}{N}\sum_{i}^{N} | \hat{x}_i - x_i |$$

Mean square error (MSE)

$$MSE = \frac{1}{N}\sum_{i}^{N} (\hat{x}_i - x_i)^2$$

Nash-Sutcliffe efficiency (NSE)

$$NSE = 1.0 - \frac{\sum_{i=1}^{N} (\hat{x}_i - x_i)^2}{\sum_{i=1}^{N} (x_i - \overline{x_i})^2}$$

At present, the regression correlation coefficient, R^2, a measure of the proportion of variability explained by, or due to the regression (linear relationship) in a sample of paired data, is not proposed to be calculated as part of the whole hydrograph performance indicator module. This parameter is commonly used by both transfer function and lumped conceptual rainfall-runoff module calibration environments. It is a number between zero and one, a value close to zero suggesting a poor model. A value close to unity, however, does not necessary indicate a good model, and statisticians advise against using this measure alone to assess goodness of fit.

On establishing the performance, the indicator is returned as a time series. This time series is a non-equidistant time series, labelled as a forecast historical within NFFS.

4.2.5 Peak prediction

NFFS can identify system performance in terms of prediction of multiple peaks in a given time series.

Peak accuracy in mean square error MSE_(MSE_PEAK)

$$PEAK\_MSE = \frac{1}{K}\sum_{k=1}^{K} (\hat{x}_k^{peak} - x_k^{peak})^2$$

where *K* is the number of peaks identified.

To establish the peak accuracy, the peak must be identified; logic from the NFFS Transformation Module is used. A peak needs to be independent of other peaks, and not simply the maximum value in a time window at the boundaries. Note that the peak of the estimated series does not need to fall exactly at the same time as the reference peak, but must be identified within a window.

The procedure for peak comparison is:

- Find peaks in the reference series, often measured or derived from telemetry
- Find the corresponding peaks in the estimated series – if there is no identifiable peak, use values at the times of peaks in the reference series
- Determine performance.

It should be investigated with the NFFS developers whether it is possible for a user to review the calculated pairs of peaks and make any manual intervention necessary before calculating the statistics.

4.2.6 Volume prediction

One important aspect of the system performance is its ability correctly to simulate the volume of water in the hydrological network. This can be assessed using the following measure of volume error:

Volume error measure PERC_VOLUME

$$PERC\_VOLUME = 100 \times \frac{\sum_{i=2}^{N} (\hat{x}_i + \hat{x}_{i-1})}{\sum_{i=2}^{N} (x_i + x_{i-1})}$$

4.2.7 Lead-time accuracy

Performance of forecasts within NFFS is assessed on the basis of 'lead-time accuracy'. This is done by comparing the forecast lead-time value against the observed value at the same time, received subsequent to the forecast. For each lead-time, this value is assessed over a given number of forecasts. An option in the configuration of the module determines if the module identifies performance of approved forecasts only, or of all forecasts.

Performance is assessed over all forecasts available for a given period of time, e.g. over a week or month, but at present no longer than one month. Lead-time accuracy is evaluated using MSE, MAE or BIAS.

Lead-time accuracy in mean square error LEAD_MSE

$$LEAD\_MSE^I = \frac{1}{J} \sum_{j=1}^{J} (\hat{x}_j^I - x_j^I)^2$$

Lead-time accuracy in bias LEAD_BIAS

$$LEAD\_BIAS^{I} = \frac{1}{J}\sum_{j=1}^{J} \hat{x}_j^{I} - x_j^{I}$$

Lead-time accuracy in mean absolute error LEAD_MAE

$$LEAD\_MAE^{I} = \frac{1}{J}\sum_{j=1}^{J} | \hat{x}_j^{I} - x_j^{I} |$$

where $LEAD\_\#\#\#^{\tau}$ is the lead-time accuracy at time τ, J is the number of forecasts considered, x_j^{I} is the reference value at time τ and $\hat{x}_j^{I}$ is the estimated value at time τ.

The results of the evaluation are written as a time series (simulated forecasting) with, as a reference time, the T0 of the evaluation run and a time stamp for each τ.

On selecting reference values x_j^{I}, these may not yet be available (should this be the case then the number of forecasts considered (*J*) is reduced accordingly). If less than the configured number is considered, then a WARN message indicating how many of the expected number were actually used.

4.2.8 Threshold crossing analysis

An important indicator of performance is the timing of predicted threshold event crossings. Again this can be evaluated over a number of forecasts. To evaluate this, the threshold crossings in the indicator and the reference series, usually from telemetry, are considered. For each pair of matching thresholds the time between the two is evaluated, and expressed either as a time bias (T_BIAS) or a time absolute error (T_MAE).

Threshold time bias T_BIAS

$$T\_BIAS = \frac{1}{J}\sum_{j=1}^{J} \hat{T}_j - T_j$$

Threshold time absolute error T_MAE

$$T\_MAE = \frac{1}{J}\sum_{j=1}^{J} | \hat{T}_j - T_j |$$

where T_j is the time of the threshold in the reference series, $\hat{T}_j$ is the time of the threshold in the estimated series.

The results of the evaluation are written as a (simulated historical) time series to the NFFS data store.

It is important to note that these statistics do not deal with the case where the indicator series did not exhibit a threshold crossing seen within the reference series. A form of penalty would have to be developed to make this measure more meaningful in this case.

In the general case, there may not be obvious pairs of matching threshold crossings, which could invalidate certain statistics. This situation should be clearly flagged to the user by the software, and strengthens the need for incorporating a facility to enable manual intervention in the selection of matching threshold crossings.

4.3 Additional performance measures needed for coastal flood forecasting

4.3.1 General approach

The content of Section 4.3 was prompted by the needs of coastal flood forecasting, and in particular how they might differ from those of fluvial flood forecasting. In principle, coastal flood forecasting can be considered in the same terms as fluvial flood forecasting. Time series of coastal variables are forecast, with periodic storms, crossings of thresholds, durations above thresholds and peak values. Therefore, as far as practical, the same types of performance measures can be used for coastal as for fluvial.

The contingency table, the reliability parameter and the unavailability of measurements (discussed in the remainder of Section 4.3) are points not exclusively relevant to coastal flood forecasting, but they are points not at present being addressed within NFFS. Even if less useful in the context of fluvial flood forecasting, any solutions to these points developed for the coastal situation would be meaningful, and could also be adopted for consistency in the fluvial situation. This would be convenient if both are to be implemented within NFFS and both are to be run within the Environment Agency's regional flood forecasting offices.

4.3.2 The contingency table

The contingency table is a convenient tool that can be driven either by observations or by measurements. It has been adopted in slightly different forms for the Met Office wave model, STFS (Table 5, Section 2.6.2), river level forecasts in Midlands Region (Table 4, Section 2.4.2), and for the overtopping rate predictions in the NW Region (Table 3, Section 2.4.1) and at Samphire Hoe, Kent. Its particular advantage for coastal flood forecasting is that where there is no instrumental measurement, it can be simplified to a basic binary system in which a predicted threshold is compared with the perceived risk of flooding assessed by an observer. This would fit in with the information available from the approach used in the NW Region, of sending an observer whenever predicted or actual sea level, overtopping rate or overtopping volume exceeds a threshold.

The format of the results summary proposed by Lukey (2003, see Table 3) involves the simplest possible comparison of an alert level being reached or not reached over a series of measured and/or forecast events. In practice, there might be several of these tables, representing different locations, different alert levels and possibly different flood risk variables.

Met Office (2004, see Table 5) includes an additional 'near-miss' event category in which the relevant threshold predicted to be exceeded by a forecasting model is not quite reached in reality. This seems a desirable enhancement to the contingency table, but one that might be difficult to define for observed (as opposed to measured) conditions. An alternative approach is already in use in Midlands Region (Cross, 2004b, see Table 3) involving a comparison of different physically meaningful threshold levels (either measured and/or forecast) in the same contingency table, e.g. flood watch, flood warning and severe flood warning.

In future, if probabilistic forecasting becomes common, it might be desirable to incorporate into the analysis the forecast probability of different alert conditions being exceeded.

4.3.3 Reliability

The term *Reliability* has been used extensively but as yet no firm definition has been given either in this report or in NFFS. To some extent it is implicit in the 'skill scores' associated with contingency tables, but the following explicit definition is suggested.

Reliability R%

$R = 100\% [1 - (Number\_of\_forecasting\_failures)/(Total\_number\_of\_forecasts)]$

where Number_of_forecasting_failures counts the number of occasions when a forecast could not be made for whatever reason, e.g. hardware or power failure, lack of input data, lack of forecasting staff, or weather conditions outside the range of validity of forecasting models, causing obviously erroneous results;

and Total_number_of_forecasts is the total number of forecasts that would have been made over the same period of time if no forecasting failure had occurred.

It is suggested that R% be calculated for each forecasting location and for different lead-times of interest. In addition to evaluating R% for all forecasts (regardless of magnitude), it may also be helpful to calculate separate values for critical forecasts, say within plus or minus two days of the flood watch threshold being exceeded. R% is a high level parameter, saying nothing about accuracy (beyond the fact that forecasting models behave broadly as intended), but suitable for dissemination to the public as a measure of the proportion of time that the forecasting service was available.

4.3.4 Particular differences from fluvial flood forecasting

Objective coastal measurements are sparse and so most of the automated performance measures proposed for fluvial use would be impractical for most coastal flood forecasting variables.

The timing of the peak value (and the uncertainty about lead-time before the peak) may be less relevant for coastal flood forecasting, depending on the variables being considered, as the only times of interest are near to high tides, whose timings are predictable.

Which is the key forecast variable (e.g. overtopping rate) may be less clear than for fluvial flood forecasting; also, its value may be far more uncertain. Whilst waves and water levels could be treated in a similar way to fluvial variables, and be tested in terms of centimetre accuracy, overtopping rate predictions and initial estimation of thresholds of interest aim only for order of magnitude accuracy.

The exact geographical boundary between 'river' and 'coast' is often unclear, but for the purpose of deciding which type of performance measures to use in estuary flood forecasting, it is probably best to draw a distinct boundary within each estuary of interest. Outer estuaries, if potential flood risk due to wave impacts dominates that due to fluvial effects, would be designated as 'coast'. Other estuary areas would be designated as 'river' and take the performance measures associated with still water levels (i.e. without waves).

4.3.5 Availability of measurements and observations

About twenty-five A Class tide gauges around England and Wales provide a continuous feed of data via the national tide gauge network. Similarly WaveNet provides real-time access to a smaller number of wave recorders, although at present most of them are offshore, outside the range of the coastal models that might be used by the Environment Agency. There are many more water level gauges owned and operated by the Environment Agency and harbour authorities, some of which are close enough to the sea to record levels representative of conditions at the coast. Additional sea level data from these gauges could be made available in real-time via local-area networks. Where available in the nearshore zone or estuaries, wave and tide gauges can be used for continuous monitoring of the performance of forecasting models, but typically there will only be two or three wave gauges and a small number tide gauges within each Environment Agency region.

It would be useful to collect data on overtopping rate and flood extent, for calibration and validation purposes, even if not for real-time use. Overtopping rate estimates could be obtained by site visits during forecast and/or actual alerts, but it would also be useful to set up CCTV at a number of vulnerable locations. Although few existing cameras would have been sited to monitor sea defences, it would be sensible to explore the possibility of sharing existing camera coverage with local authorities, if legal issues about privacy could be resolved. It would provide a valuable source of data if flood extent could routinely be recorded during actual coastal flooding events, either by site visit or photography.

It is difficult, even for an experienced observer, to convert observations or images of overtopping, into overtopping rates, and it may be more practical to use subjective descriptions of rates instead. Figures 5-7 are three photographs of overtopping occurring at Samphire Hoe, near Dover. The representative mean overtopping rates shown in these photographs are classified by HR Wallingford as follows:

- Figure 5: Normal public hazard, mean discharge about 0.03 litres per second per metre. These conditions are considered to be just tolerable for ordinary members of the public.

- Figure 6: Safe only for skilled operatives who are aware of the potential dangers, mean discharge about 0.3 litres per second per metre. Peak individual volumes one hundred times larger than the mean discharge rate should be anticipated.

- Figure 7: Extreme danger and potentially fatal. The figure shows a peak individual overtopping volume in excess of 1.5 cubic metres per second per metre. These peak events are sporadic and the storm may sustain mean discharges rates of 3.0 litres per second per metre for long periods.

Until the Environment Agency gains more experience in observation of overtopping, these photographs could be used to provide an approximate 'calibration' of CCTV monitoring images.

Figure 5 Example of mean overtopping rate of about 0.03 l/m/s at Samphire Hoe (photo HR Wallingford)

Figure 6 Example of overtopping rate of about 0.3 l/m/s at Samphire Hoe (photo HR Wallingford)

Figure 7 Example of peak overtopping occurring during mean overtopping rate of about 3.0 l/m/s at Samphire Hoe (photo HR Wallingford)

4.4 Operation of performance measures

4.4.1 When, where and how will the measures be applied

The best option would be to incorporate any required performance measures into NFFS, and to apply them as consistently as practical across different regions and different variables within fluvial and coastal flood forecasting. They should, of course, take account of any false alarms where no significant flood risk occurred. NFFS will be run from the Environment Agency's eight regional forecasting offices.

All of the measures of time series analysis already intended to be implemented in NFFS will be produced routinely on every forecast run, looking back over fixed period(s) of time prior to the forecast.

Mostly the same measures can also be produced to focus on user-specified periods of time of particular interest or importance. These calculations might be made after events (whether forecast and/or actual), over periods of time set to focus on a whole hydrograph, and/or to measure performance during the occurrence of a single peak or a group of peaks.

The summary performance measures, often described as 'high level', will be produced periodically, depending on reporting requirements. The contingency table (and any associated skill scores) and the reliability parameter are of this type.

Since the generation of performance measurement reports in NFFS is automatic, it should be relatively simple to configure regional systems to measure the effect of and establish the dependency on sub-processes within a forecasting system such as:

- Accuracy of forecast rainfall or snowfall
- Accuracy of forecast surge
- Efficacy of module data set calibration
- Accuracy of particular forecasting modules, or types of forecasting module at given lead-times in advance of a peak or threshold crossing
- Dependency on error correction at particular or multiple locations
- Dependency on operation of moveable structures.

It is hoped that new performance measures recommended in this report will be incorporated into the next 'Phase 3' version of NFFS, and that refinements will be made after an initial trial period, before adoption across other regions. To facilitate this, it will be necessary for NFFS to be able to accept observational data (e.g. overtopping rate or flood extent) and to be able to process up to one year of data to calculate the periodic summary statistics.

The part of England and Wales currently best served with measurements of coastal flood risk source variables is the Channel Coastal Observatory (CCO) area between Bournemouth and Folkestone. Measurements are available in near real-time through the CCO web-site at http://www.channelcoast.org. The observatory includes several wave gauges at coastal locations, several tide gauges not available as part of the national tide gauge network, and a smaller number of meteorological stations. These data could be used both for real-time updating of coastal forecasting models and for subsequent evaluation of performance measures.

It may not be practical to introduce coastal flood forecasting into the initial phase of NFFS at this late stage, but as Southern Region is involved in this phase, it would be advantageous to do so. Continued development of coastal flood forecasting, and implementation and testing of associated performance measures, would be helped by Southern Region being involved in this topic during the initial phase, to take advantage of the CCO measurements. Initial validation and performance measurement trials for coastal flood forecasting services could be concentrated here, and in North West Region where coastal flood forecasting is well established. Experience gained in areas well served by measurements could subsequently be applied in other regions, in terms of data acquisition, wave and tide model calibrations, and choice of forecast thresholds.

4.4.2 Issues

The following issues regarding operation of performance measures were identified by the project Board.

- The time resolution of system runs may not be adequate at the time of peak or threshold crossing. It may be desirable to initiate and process additional runs, automatically or manually, but this may impose an unacceptable loading on the real-time system.
- How easy or costly is it to include additional performance statistics in NFFS and when could this be done?

- It was felt by the Project Board that there was a need for a graphical representation of forecasting performance for each location, such as a bulls-eye plot for each predetermined lead-time. An example is given in Appendix 2.
- It was not the view of the Project Board that there was a need either for a graphical representation of forecast convergence as a function of lead-time, or statistical measurement of the predicted duration of flooding above a given threshold.
- Should the 'whole hydrograph' performance measures be constrained to analysis of values above a configurable threshold?
- It would be helpful, both for trial and comparison purposes for the summary performance statistics, to recover and process forecast data from previous years. This is intended to be possible within NFFS, but it is not yet clear how much earlier forecast data will be easily recoverable.

5. RECOMMENDATIONS FOR USE OF PERFORMANCE MEASURES FOR FLOOD FORECASTING

Chapter 5 begins by summarising the alternative options for performance measures, and the relationship of those options to existing methods to be implemented within NFFS. It goes on to recommend one of those options, and to outline a programme for implementation and testing within the Environment Agency Regions. The recommendation is, initially, to accept all the performance measures already being developed within NFFS, plus several additional parameters. These include the existing R^2 correlation coefficient parameter, a new timing error parameter, a new reliability parameter, and a summary contingency table. Chapter 5 concludes by outlining further R&D required to facilitate performance evaluation.

5.1 Discussion of options for real-time performance measures

1. Existing NFFS options only

An advantage of this approach is that no additional work is necessary. A further advantage is that whole hydrograph methods can establish model performance over the full range of possible flows, which may have value for low flow modelling.

A possible disadvantage is that 'whole hydrograph' performance measures do not focus on model performance at high flows, which is generally of more significance.

It is suggested that the lead-time accuracy time series should be developed using a selection or 'ensemble' of forecasts whose number should be configurable, but care would have to be taken to ensure that this number is not so large that the measure loses focus.

The unit of these measures is the same as that of the two series being compared – usually water level in units of metres.

These proposed measures are not used in any of the Agency's existing regional forecasting systems and may initially be unfamiliar to forecasters, notwithstanding any training that they will receive.

For the measures to be meaningful, the granularity in time of the forecasts must be similar to the lead-time for which measures are required. For example, if the forecasts are made only once a day and a given peak or threshold crossing occurs roughly half-way between the forecasts, then the lead-time accuracy time series may not be helpful in assessing how well the modules and data sets performed in this case.

2. Existing NFFS options, plus contingency table and regression correlation coefficient R^2

One of the measures used in various forms at present is the contingency table, indicating simply whether thresholds were exceeded or not, and thus whether appropriate warnings could be issued. The contingency table is a useful broad binary assessment of system performance, especially if associated with particular lead-times as used in Met Office (2004) or Cross (2004b). The measure would ideally incorporate the timeliness, via a configurable time period, of the forecast threshold crossings because, for locations where error modelling is applied, the particular threshold crossing should always be

correctly predicted if the lead-time is sufficiently short. If error correction were not applied at a particular location, however, the measure would be of some value even at very short lead-times.

The feeling of the Project Board was that the 'near-miss' category used by Met Office (2004) would not be helpful, as a number of different thresholds are already used by flood forecasters. However, a few 'skill scores' summarising the information in contingency tables still further, are helpful in providing comparisons between regions or between periods of time. 'False Alarm Rate' and 'Probability of Detection' suggested by Lukey (2003, see Table 3) and 'Total Timeliness' and 'Total Accuracy' used by Cross (2004b, see Table 4) seem to be the most useful and easily understood skill scores.

The advantage of adding the R^2 measure is that it is often used in rainfall-runoff modules' native calibration environment such as TSCAL, owned and marketed by CEH Wallingford. Whilst this statistic is not an ideal measure, as discussed in Section 3.1.9, it may provide values that are familiar to forecasting module data set developers.

3. Existing NFFS options, plus contingency table and regression correlation coefficient R^2, plus separated n hour lead-time statistics for peak magnitude error and time of peak error

The advantage of separated lead-time statistics for identified peaks only is that the user should be familiar with the measures presented even if the system was not actually run at the times for which the measures are calculated. It is relatively simple to present these additional measures on a 'bulls-eye plot' (see Appendix 2) for a given lead-time before the observed peak, in which the y-axis would be relative magnitude error in metres or millimetres. This type of plot could be used both for the current event and for a selection of historical events, to indicate forecast performance at specific forecasting points.

These performance measures represent the optimal possible performance of the underlying modules and data sets within a given system, but in reality the system will not be run in real-time at precisely the times where the performance measures are calculated. This introduces an additional error source into the forecast which is dependent on the implementation of the forecasting system and specifically the granularity in time of the system simulations.

4. Existing NFFS options, plus contingency table and regression correlation coefficient R^2, plus separated n hour lead-time statistics for peak magnitude error and time of peak error, plus the same for each threshold crossing

This option introduces the same approach as for the peak prediction for the threshold crossings. Once a threshold at a forecasting point has been crossed, the system should be able to calculate the n hour lead-time statistics for that particular threshold crossing at that forecasting point, thus calculating once again the optimal performance statistics for the modules and data sets.

A factor to be borne in mind for this approach is that a large amount of information would be presented to the forecaster which may be difficult to assimilate in real-time. Furthermore, if n hour lead-time statistics are required, for N lead-times, for T threshold

crossings, and for K forecasting points, the number of runs formally required is the product of N, T and K. It may not be feasible to carry out this many runs during or even subsequent to an event due to time constraints.

In reality it may be necessary, especially during an event, to investigate the feasibility of interpolating performance measures for a given set of system simulations at particular times which may not be equidistant, but this would introduce an additional potential source of error.

5. Existing NFFS options, plus contingency table and regression correlation coefficient R^2 with configurable thresholds, plus separated n hour lead-time statistics for peak magnitude error and time of peak error, plus the same for each threshold crossing

This is the same as Option 4 but with the 'whole hydrograph' measures applied only to flows or levels above (or below for low flow simulations) a configurable value. This option was strongly supported by the Project Board.

6. As Options 1-5 but with a restricted subset of existing NFFS options

The proposed set of NFFS performance measures is comprehensive (although some of these measures ultimately may not be required). Before such decisions can be made, however, it would be sensible to make an assessment of their relative values on the first three NFFS regional system implementations.

Table 9 summarises the features contained in each of the above options.

Table 9 Summary of performance measures included in the implementation options

Performance measure		Option					
		1	2	3	4	5	6
NFFS		✓	✓	✓	✓	✓	½?
Contingency Table		×	✓	✓	✓	✓	✓
R^2		×	✓	✓	✓	✓	✓
R^2 with configurable thresholds		×	×	×	×	✓	✓
Separated n hour lead-time statistic for…	Peak magnitude error	×	×	✓	✓	✓	✓
	Time of peak error	×	×	✓	✓	✓	✓
	Threshold crossing error	×	×	×	✓	✓	✓

Although *reliability* is frequently mentioned in the context of performance measurement, there is currently no agreed measure for reliability. It is suggested that the definition given in Section 4.3.3 be considered for possible adoption. It could be applied separately to different forecasting locations and different lead-times, and could be adapted to focus on critical forecasts.

In future it may be worth considering measures which do not involve squared difference terms. One such measure might be the modified Nash-Sutcliffe Efficiency statistic NSE_1 defined below

$$NSE_1 = 1.0 - \frac{\sum_{i=1}^{N} | O_i - P_i |}{\sum_{i=1}^{N} | O_i - \overline{O_i} |}$$

where O and P denote observed and predicted values respectively.

As discussed in Section 2.6.1, another measure that could be calculated is the timing error. A cross-correlation analysis could be carried out between the predicted and measured data for different time lags, to provide an indication of whether the predictions are systematically early or late.

5.2 Recommended option for performance measures

Option 5 of the above measures represents a sensible initial option for use within NFFS. It uses all of the proposed measures by Delft Hydraulics and Tessella, plus additional measures focussed explicitly on peak timing and magnitude prediction. Specified lead-times will focus on clear objectives such as the satisfaction of the Agency's customer charter requirements, implementation of Major Incident Plans, minimisation of economic loss and time required to construct temporary or demountable defences. All of these have designated lead-times.

As experience is gathered in the use of the performance measures, certain of them may not add value to the assessment of performance and may be dropped as a consequence, leading to Option 6 in the longer term.

The 'contingency table' approach provides a summary assessment of system performance, taking into account lead-times, and it is recommended that this is also included as a high level measure of system performance in the automatic periodic NFFS reports. Further recommendations for the future, outlined in Item 7 of Section 5.1, are for the implementation of a reliability parameter, a timing error calculation and calculation of NSE_1.

It is also recommended that statistics are kept describing the differences in forecast performance between actual forecasts made during an event and post-event hindcast performance (where actual times of peaks and threshold crossings are known). This would assist in distinguishing between sources of error from the forecasting model, the forecasting systems use of the model, and from the input forecasts.

5.3 The particular parameters to be performance measured

The processes and parameters listed in Table 10 (reproduced here from the four-level classification introduced in Table 8) fall within the scope of this report. Here, they are listed with an added fifth-level note on the general type of performance measures to use.

Table 10 Summary of the types of performance measures to apply to different physical processes and parameters involved in flood forecasting

Flood forecasting
River
River flow
– Peak flow over a fixed period of time, e.g. 1 hour
(NFFS performance measures)
– Time of peak flow
(NFFS performance measures)
– Shape of the flow curve over time, e.g. 20 hours
(NFFS performance measures)
River level
– Peak level near the time of an event
(NFFS performance measures)
– Time of peak level
(NFFS performance measures)
– Occurrence of threshold crossings
(NFFS performance measures)
– Time of threshold crossings
(NFFS performance measures)
Nearshore
Sea level
– Peak surge near the time of an event
(same as NFFS fluvial measures)
– Time of peak surge
(same as NFFS fluvial measures)
– Peak sea level near the time of an event
(same as NFFS fluvial measures)
Waves
– Peak wave height over a fixed period of time, e.g. 1 hour
(same as NFFS fluvial if measurements available) (observations and contingency table elsewhere)
– Peak wave period over a fixed period of time, e.g. 1 hour
(same as NFFS fluvial if measurements available) (observations and contingency table elsewhere)
– Time of peak wave height
(same as NFFS fluvial if measurements available) (observations and contingency table elsewhere)
Defences
Overtopping
– Peak overtopping rate over a fixed period of time, e.g. 1 hour
(observations and contingency table)
– Volume of overtopping during an event, e.g. over high tide
(observations and contingency table)
– Time of peak overtopping (fluvial only)
(equivalent to time of peak river level)

Threshold levels
River
River level
– Occurrence of different threshold levels
(measurements and contingency table)
Nearshore
Overtopping rate (and/or other, e.g. sea level and wind)
– Occurrence of different threshold levels
(observations and contingency table)

5.4 Outline programme for implementation within the Environment Agency

There are two potential pathways for implementation of the measures recommended in this report. The first and primary application is NFFS, both during an event, in post-event analysis, and to present longer term forecasting system performance.

The second potential application is within calibration environments associated with forecasting modules such as PDM, KW, ISIS and MIKE11. In particular, it would be useful to include these measures in the TSCAL environment which is the only current means by which PDM or Thames Conceptual Model (TCM) data sets can be developed for use within the NFFS. In parallel to the development of NFFS, discussions should be held with model developers to investigate the feasibility of incorporating a revised set of performance measures into native calibration environments. This would have the benefit of allowing comparison between the performance of offline and online model data sets over multiple performance measures.

As for NFFS itself, another development phase is planned to implement the additional requirements of the five other regions. It would seem appropriate initially to discuss the recommendations within this report with both the NFFS team and Delft Hydraulics, prior to an amendment to the current proposed development programme. The next 'Phase 3' development is scheduled to begin in April 2005.

A minimum duration to assess the current measures would be the 4-6 months covering a complete flood season. The assessment should investigate performance for at least six forecasting points representative of different flooding timescales and forecasting modules within each of the initial three regions. Ideally, the review at each location would consider data from at least three events.

The current NFFS measures (see Section 4.2) will be implemented during the Site Acceptance Testing period. This should provide statistics which can be assessed prior to the proposed Phase 3 development. A further review of performance measures should be applied over a period of 12 months, after the Phase 3 developments have been implemented.

5.5 Further research and development

5.5.1 Introduction and outline programme

It is assumed that continued development and operational testing of NFFS, and implementation of the recommendations of the present project, will be funded by the Environment Agency. This section outlines three additional research topics (5.5.2-5.5.4) related to performance measurement and improvement, and one (5.5.5) related to valuation of flood forecasting. These were originally written near to the start of the present project, in about May 2004.

The ideas outlined in Sections 5.5.2-5.5.4 were subsequently collated by HR Wallingford, in Environment Agency Short Form A format, together with a further idea for a demonstration coastal flood forecasting system at a single site. This composite outline proposal was put forward by Bob Hatton at the Flood Forecasting Theme meeting on 20 July 2004. If accepted as it stands, the work would be done over a two-year period beginning in about April 2005.

The ideas described in Section 5.5.2 have subsequently been adopted as part of the work intended to be taken forward over the coming few years within the flood forecasting elements of the Flood Risk Management Research Consortium. It would therefore not be appropriate to allocate new funding to this topic, at least not without reference to the work within FRMRC.

The topic described in Section 5.5.5 is rather different, in that it focuses on the value of flood forecasting and warning relative to the value of other aspects of flood management such as flood defences. This is envisaged as a one-year stand-alone project, providing information to the Environment Agency to assist in setting an appropriate level of investment in the flood forecasting and warning service. Although costs and benefits may change following the implementation of NFFS, this topic was regarded as a priority by the Project Board, and there is no scientific reason why it could not begin immediately.

5.5.2 Uncertainty propagation: A framework for real-time forecasting

Flood forecasting modelling solutions are constructed of a series of models, some coupled internally (e.g. wave, current and wind models) and some coupled externally (e.g. discrete models of nearshore waves and defence overtopping), often without clearly defined procedures for the transfer of data and information. Traditionally each model has been treated as essentially deterministic, providing a single forecast to the next model in the system. However, current interest is focused on identification of the uncertainty associated with an individual model output and the propagation of this uncertainty forward in the coupled model chain through the source, pathway and receptor variables. This research topic is focused on uncertainty propagation through complex real-time modelling capabilities and its assessment at the interfaces. The results should be aimed at developing a unified approach to generating estimates of uncertainty in complex coupled modelling in fluvial, estuarial and coastal environments. (Linked to this topic, and to some extent the next topic, is the concept of ensemble modelling used in some meteorological applications to quantify the sensitivity of forecasts to assumed uncertainties in input values.)

5.5.3 A risk based flood forecasting modelling framework

Flood forecasting will benefit from the recently developed NFFS. However, the process of providing a flood forecast is a generic one and only the complexity of the source, pathway and receptor models vary. This generic process would be developed within the context of an open architecture framework to provide a specific Modelling Decision Support Framework (MDSF) for Forecasting. This would include a common approach to determining risk, presenting results, accounting for defences and accessing defence data, and propagating and displaying uncertainty. The approach developed is likely to build upon the research completed under the RASP project: Risk Assessment of flood and coastal defence systems for Strategic Planning. This would need to be modified to account for a forecast loading condition as well as the approaches to receptor risks included within the MDSF developed to support Shoreline Management Plans.

5.5.4 Data assimilation: Real-time updating of flood forecasts

Real-time flood forecasting systems allow the flow of new data from a variety of sources during the progress of a flood event. Whilst this is reasonably well developed for fluvial forecasting, at present the assimilation of data into coastal flood forecast models is difficult and largely restricted to models run by the Met Office. However, information gathered by the Agency during a storm – including automated as well as observational measurements – provide a real opportunity to improve forecasts locally. The proposed research would need to use different mathematical structures involving various updating techniques to ensure the information content from data is maximised across source, pathway and receptor models. In particular this research will need to build upon existing research knowledge and tailor the approaches to the forecast needs of the Agency and the data they have available. If successful, the research will maximise the use of all data within the system models and build a real-time "learning capability" into forecast systems.

5.5.5 The relative investment performance of flood defences and flood forecasting

Construction and maintenance of fluvial and coastal defences each have their own budgets and priorities, as do fluvial and coastal flood forecasts. All are expected to provide value for money. Methods for estimating the investment return on defences are well established, and are applied and regulated consistently across England and Wales. It would be interesting to attempt a similarly rigorous approach to the benefits and costs of flood forecasting in different types of flood risk area and to the overall national service. This could be applied separately to initial costs and to running costs, and could perhaps consider a range of alternative levels of forecasting service from minimum to optional extras such as site-specific measurements for real-time assimilation. This would allow the relative benefits of flood defences and flood forecasting to be assessed on a fair and consistent basis, in turn helping to guide the investment strategy for flood forecasting in the future (e.g. increase investment, withdraw service from some areas, no change).

6. REFERENCES AND OTHER INFORMATION RECEIVED

Abernethy R, Tozer N and Hulse D (2004). Evaluation of an operational nearshore wave forecasting system against measured data. TBA.

Behan R (2004). R&D flood forecasting performance measures. Personal email communication on 22/06/04, giving examples of performance monitoring in the SW Region.

Cross R (2004a). Feb 2004 forecasting model analysis. Personal email communication on 29/04/04, giving an example of performance monitoring at Bewdley, carried out by the Midlands Region.

Cross R (2004b). Leadtime analysis for Feb 04.xls. Personal email communication on 01/11/04, giving an example of a performance summary for the whole of the River Severn, carried out by the Midlands Region.

Defra (2004). Flood and coastal defence project appraisal guidance: FDCPAG6: Performance evaluation. Defra publication to be issued soon.

Defra / Environment Agency (2003a). Best practice in coastal flood forecasting: Technical report. Defra / Environment Agency R&D Technical Report FD2206/TR1. Issued for use by the Environment Agency as HR Wallingford Report TR 132.

Defra / Environment Agency (2003b). Guide to best practice in coastal flood forecasting. Issued for use by the Environment Agency as HR Wallingford Report SR 618, including CD version with many digital links to FD2206/TR1 (TR 132).

Delft Hydraulics and Tessella (2004). National flood forecasting system: Performance indicator module design. Delft Hydraulics report to the Environment Agency.

Environment Agency (1998). Tidal flood forecasting project report. Environment Agency, October 1998.

Environment Agency (2000a). Flood warning investment strategy 2000/01 to 2009/10, Issue 1.1, June 2000.

Environment Agency (2000b). Concerted Action for Flood Forecasting and Warning: Workshop report and proposals for action – Technical Report W220.

Environment Agency (2001a). Flood warning service strategy for England and Wales.

Environment Agency (2001b). Reducing flood risk: A framework for change. Environment Agency, July 2001.

Environment Agency (2001c). Customer Charter.

Environment Agency (2001d). Report on performance of existing flood forecasting models. Internal Environment Agency report by the NE Region Forecast Improvements Project Team.

Environment Agency (2004a). Work instruction: Flood warning investment strategy performance measures. Environment Agency Management System Document No 359_03, Version 2, January 2004.

Environment Agency (2004b). Flood warning investment strategy 2002/03 – 2112/12 interim performance measurement tool (IPMT) procedures manual. Environment Agency, April 2004.

Environment Agency / Defra (2002). Forecasting extreme water levels in estuaries for flood warning: Stage 2 – Review of external forecasts and numerical modelling techniques. Environment Agency / Defra R&D Project Record W5/010/2.

Environment Agency / Defra (2003). Fluvial flood forecasting for flood warning: Real time modelling. Environment Agency / Defra R&D Technical Report W5C-013/5/TR, March 2003.

Environment Agency / Defra (2004). Protocols for minimum standards in modelling (Flood warning management system Phase 2a). Environment Agency / Defra R&D Technical Report W5C-021/TR.

FloodWorks (2004). Wallingford Software report.

Green H (2004). R&D flood forecasting performance measures. Personal email communication on 23/06/04, giving a description of forecasting model validation in the Thames Region.

Khatibi R, Stokes R, Ogunyoye F, Solheim I and Jackson D (2003). Research issues on warning lead-time and synergy in flood mitigation measures. International Journal of River Basin management, Volume 1, No 4, pp331-346.

Khatibi R, Jackson D, Harrison T, Price D and Haggett C (2004). Definition of best practice in flood forecasting. Special edition of HESS.

Lalbeharry R (2001). Evaluation of the CMC regional wave forecasting system against buoy data. Atmosphere-Ocean 40 (1) 2001, pp1-20.

Legates D and McCabe G (1999). Evaluation of the CMC regional wave forecasting system against buoy data. Water Resources Research 2001, Volume 35, No 1, pp233-241.

Lukey B (2003). Proposal for defining required flood warning accuracy. Internal Environment Agency discussion document dated 24/11/2003.

Lukey B (2004). Performance measures for flood forecasting. Personal email communication on 22/04/04, describing performance measures being developed for the NW Region TRITON forecasting system.

Met Office (2004). Storm Tide Forecasting Service: Annual report to Defra for the 2003/4 operational season. Met Office report, Crown.

Ministry of Agriculture Fisheries and Food, Environment Agency, Met Office and Proudman Oceanographic Laboratory (1998). Tidal flood forecasting joint action plan. October 1998.

Modarres M (1993). What every engineer should know about reliability and risk analysis. Marcel Dekker Inc, New York.

Office of Naval Research (2002). The ONR Test Bed for coastal and oceanic wave models. Manuals Version 2, December 2002.

World Meteorological Organisation (1998). Marine meteorology and related oceanographic activities: Report No 36, Handbook of offshore forecasting services: Section VI Forecast verification. WMO Report WMO/TD-NO 850, http://info.ogp.org.uk/metocean/OWP/section_6.html.

APPENDICES

Appendix 1

Use of fluvial flood forecasting performance measures within the Environment Agency regions

A1.1 Anglian

No information provided to date.

A1.2 Wales

It is understood that Welsh Region undertakes general post-event analyses but no details of these have been provided.

A1.3 Thames

Thames Region forecasters undertake post-event analyses of their rainfall-runoff model performance, and of the performance of the forecasts in their GeoGUI system.

The Cascade system has recently been enhanced to generate historic predicted time series when required, which greatly assists the procedure to assess the RFFS model performance. Generally, the output of the RFFS system focuses on the R^2 regression correlation coefficient over the whole hydrograph, which compares the variance of the observed and simulated time series.

The post-event analysis undertaken using GeoGUI is similar to that described for North West region whereby the T hour lead-time predictions can be shown on an observed time series plot at a specified forecasting point. The post-event analysis also presents the mean absolute error (MAE) and maximum error for specified lead-times in advance of peak observed values.

Performance measures are not automatically produced by GeoGUI but the required information can be extracted relatively easily.

A1.4 North East

A document describing post-event analysis of the current RFFS performance for the 2000 flood event was provided by North East Region.

The approach used in this document is to analyse the prediction of peak magnitude and peak timing for model runs for a single location undertaken at several times relative to the observed peak level. Information on threshold crossings is also tabulated in the report and this information is presented as 'not observed' when the threshold was not predicted to be crossed by the model and as a timing error in hours where it was.

Pros

- Many runs undertaken enable presentation of the forecast evolution as a function of lead-time.

- Peak prediction, timing of peak prediction and threshold timings are assessed for different lead-times.

Cons

- It is difficult to compare with other regions' approaches or between different sites without common timings of model simulations relative to peak or threshold crossing.
- The graphs contain much useful information, but are not easy to read.
- The threshold crossing tables are not clear.
- The magnitude error for threshold crossing is not shown explicitly.
- The importance of errors in the forecast data is not shown.

A1.5 North West

This approach focuses on the predicted level hydrographs for a particular forecasting point at a series of identified lead-times, 2, 4, 6 and 8 hours before the observed peak at the site, together with the full observed hydrograph. These results are then condensed into another graph which shows points corresponding to the four forecasts on another graph with observed level as the y-axis and time since detection in hours as the x-axis.

The approach described by Lukey (2004) also allows errors due to poor forecast values to be identified by comparison with simulations where perfect foreknowledge is presumed.

For the forecast results presented, the predicted peak varied considerably between the four simulations presented and additional simulations and one hour time intervals would have been useful to establish the evolutionary pattern of the forecast predictions as lead-times decreased to zero.

A similar approach is used with the GeoGUI post-event analysis capability, where results are presented as hydrographs at given lead-times, together with an observed peak level plot on which is presented points corresponding to forecast peak levels and timing of peak levels. A similar plot could also be presented for threshold crossing predictions rather than peak prediction.

Pros

- Fixed and definable lead-times are used which could be used for other forecasting modules and other regions. Results from WRIP and GeoGUI, two quite different forecasting systems, are presented in a common format.
- Whole hydrographs are shown which show volume and shape comparison with measured data.
- The approach could be used both for level and flow measurements and for model results comparison.
- The effects of forecast input errors are investigated by comparison with contemporary predictions during an event and with those using perfect foreknowledge.
- The graphical presentation of output is clear and understandable.

Cons

- It may not be able to establish evolution of peak and timing of peak predictions, using such a small number of model simulations.
- The x-axis of 'time since detection' seems somewhat arbitrary.
- No explicit procedure to present performance of the model in terms of threshold crossing.

A1.6 Midlands

An example from Richard Cross (2004a) was provided for the February 2004 event at Bewdley on the River Severn. This approach shows the results of multiple model runs at a time interval of 2 hours between runs from 4 days before the peak and the time of measured peak.

The two graphs presented show the evolution of predicted peak level as a function of lead-time and the evolution of predicted timing of peak level relative to lead-time

The Midlands system can also present timings of threshold crossing as a function of lead-time.

Pros

- Many model runs are undertaken which clearly allow the evolution of forecast peak levels and timings to be seen.
- Information is clearly presented.

Cons

- Many model runs are presented at times before peak which may tax the processing power of an automatic system, although this is possible as there is a considerable lead-time between the rainfall and flood peak at this location. It may not be necessary to undertake every run shown to illustrate the model's performance. A subset of 72, 48, 24, 12, 8, 6, 4 and 2hours would probably suffice in this case.
- The results are shown on two graphs rather than one, which can be difficult to cross-correlate.
- Hydrographs are not shown, so volume, shape and recession prediction cannot be assessed easily.
- Threshold crossing analysis is not presented.
- The importance of errors in the forecast input data is not shown.

A1.7 South West

The post-event analysis approach taken by South West region differs in that it focuses on whether warnings were issued correctly or not for specified events and locations. A table is produced which shows actual warnings issued as one axis and simulated warnings on the other (Behan, 2004).

A further 'model performance plot' is produced for each site which presents the extent in metres of forecast errors per event with Catchment Wetness Index and cumulative rainfall rate.

No information is currently presented for the predictions of timings of peaks or threshold crossings.

A1.8 Southern

It is understood that Southern Region undertake general post-event analyses but no details of these have been provided.

Table A1.1 Summary of performance measures in existing regional systems

Region	Automatic or manual procedure?	Peak magnitude analysis	Peak timing analysis	Threshold crossing analysis	Timing of threshold crossing analysis	Whole hydrograph analysis
North East	Manual, post-event	Yes	Yes	Yes	Yes	No
North West	Manual, post-event	Yes	Yes	No	No	Yes
Midlands	Manual, post-event	Yes	Yes	Yes	Yes	No
Thames	Manual, post-event	Yes	Yes	No	No	Yes
Southern	None	N/A	N/A	N/A	N/A	N/A
Anglian	??					
Wales	None	N/A	N/A	N/A	N/A	N/A
South West	Manual, post-event	Yes	No	Yes	No	No
NFFS	Automatic, during and post-event	No (combined)	No (combined)	Yes	Yes	Yes

Appendix 2

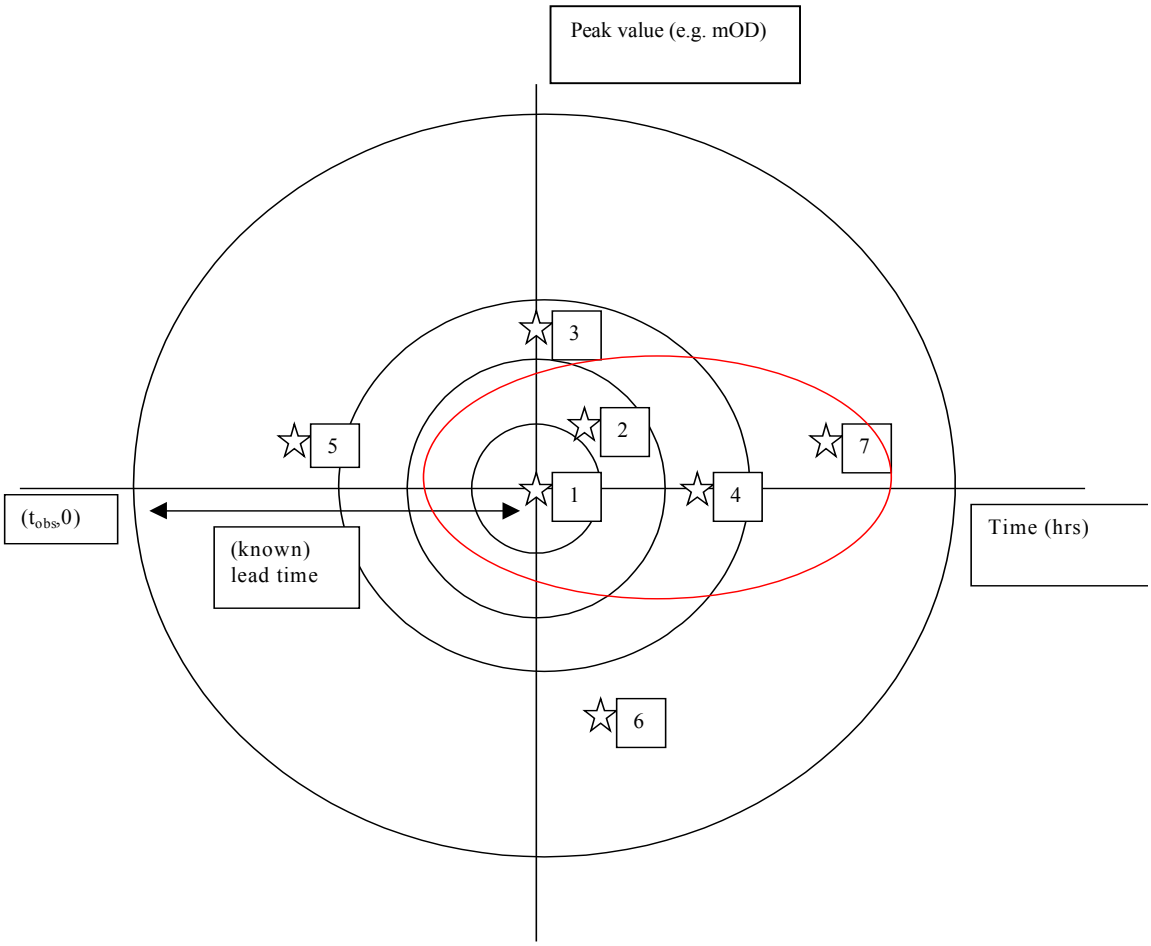
Example of a bulls-eye plot performance measure

The objective here is to show, for a given lead-time as part of a post-event analysis, how well each model (module data set) performs for a sequence of historic events (numbered 1-7).

Event 1 shows perfect prediction for peak value and time-of-peak.
Event 3 shows perfect prediction for time-of-peak but over-estimates peak value.
Event 4 shows perfect prediction for peak value but a late time-of-peak.

The combined event error is represented by the length of the line between the point and the bulls-eye. The peak value and time-of-peak errors are represented by the vertical and the horizontal distances between the event points and the respective axes.

From these, a 'deviation ellipse', shown in red on the diagram below, can be plotted. This is centred on the point defined by (average time-of-peak, average peak error). The area of the ellipse is indicative of the combined mean deviation. The lengths of the major and minor axes are indicative of the separate peak value and time-of-peak mean deviations.


Peak value (e.g. mOD)
3
2
5
7
1
4
(t_obs,0)
(known)
lead time
Time (hrs)
6