



Numerical Weather Prediction

An investigation of the ability of a storm scale configuration of the Met Office NWP model to predict flood-producing rainfall.

**Final scientific report from the storm scale modelling project.
The storm scale modelling project was half funded by Defra, project code FD2207**



Forecasting Research Technical Report No. 455
Joint Centre for Mesoscale Meteorology Report No. 150

Nigel Roberts

19th January 2005

©Crown Copyright

PRINCE2

Author: Nigel Roberts

Document Number: Version 4.0

Revision History

Revision date	Previous revision date	Summary of Changes	Changes marked
18/10/04	-	First completed draft apart from Mesoscale Model section, version 1.0	
30/11/04	18/10/04	Modifications following comments from project board, version 2.0	
10/12/04	30/11/04	Addition of Mesoscale model section 7, version 3.0	
05/01/05	10/12/04	Minor modifications to section 7, version 4.0	

Approvals

This document requires the following approvals.

Name	Signature	Title	Date of Issue	Version
Brian Golding	<i>Brian Golding</i>	Head of Forecasting Research, Met Office, Hd(FR)	19-01-05	4.0

Contents

1	Introduction to the project	5
1.1	The five stages	5
2	Why is a storm-scale model needed	7
2.1	The current state of the art in forecasting convective storms	7
2.1.1	Current advection nowcasting systems	7
2.1.2	Current operational NWP models	8
2.1.3	The Convection Diagnosis Project (CDP)	8
2.1.4	Forecaster skill and interpretation	8
2.2	The anticipated benefit of a storm-scale model	9
3	The set up of a storm-scale model	11
3.1	Future ambition and current testing	11
3.2	Collaboration with the High Resolution Trial Model (HRTM) project	11
3.3	The model configuration – Technical decisions	11
3.3.1	Domain size, domain location and horizontal resolution	12
3.3.2	Vertical resolution	13
3.3.3	Time step	13
3.3.4	Diffusion	14
3.3.5	Boundary Updating	14
3.3.6	The critical relative humidity (RH-crit)	15
3.4	The model configuration – Scientific decisions	15
3.4.1	The representation of convection	15
3.4.2	Data assimilation - Getting the best possible analysis	18
3.4.3	The representation of clouds and precipitation	20
3.5	The model configuration - representation of the land surface	21
4	Encouraging results from initial case studies	23
4.1	The Case studies	23
4.2	Case 1, 2-3 July 1999	24
4.3	Case 2, 11-12th October 2000	26
4.4	Case 3, 3rd May 2002	28
4.5	Case 4, 29th July 2002	30
4.6	Summary of the results from the case studies	32
4.6.1	Subjective performance scores	32
5	Products from a storm-scale model	34
5.1	Why are post-processed products required	34
5.2	Examples of forecast products for flood prediction	35
5.2.1	Products displayed on square areas	35
5.2.2	Products for river catchments	37
5.2.3	The use of probabilities	38
5.3	Hydrological applications	39
5.4	General comments	42
6	How to verify the model objectively	43
6.1	The verification problem	43
6.2	Verification questions	44
6.3	A verification method – the comparison of fractions	44
6.3.1	Basic decisions	44
6.3.2	Spatial scales - Computing fractions over different sized areas	44
6.3.3	A verification score to compare fractions	47
7	The skill of the operational mesoscale model analyses and forecasts during 2003	49
7.1	Outline of the investigation	49
7.2	The bias in rainfall amounts	50
7.3	Scale-selective skill of the spatial distribution of the precipitation forecasts	52
7.4	The scales over which the data assimilation operated	53
7.5	The retention of skill	54
7.6	Main conclusions	55

<u>8</u>	<u>Storm-scale model performance scores</u>	56
<u>8.1</u>	<u>Scores for 6-hourly accumulations</u>	57
<u>8.1.1</u>	<u>Fractions skill score</u>	57
<u>8.1.2</u>	<u>Brier skill scores</u>	58
<u>8.2</u>	<u>Scores for hourly accumulations</u>	59
<u>8.3</u>	<u>Implications of the results</u>	61
<u>8.3.1</u>	<u>The impact of resolution</u>	62
<u>8.3.2</u>	<u>The impact of data assimilation</u>	62
<u>8.3.3</u>	<u>Final comments</u>	63
<u>9</u>	<u>The Boscastle Flood</u>	64
<u>9.1</u>	<u>The event</u>	64
<u>9.2</u>	<u>Forecast Products</u>	65
<u>9.3</u>	<u>General comments</u>	70
<u>10</u>	<u>Conclusions</u>	71
<u>10.1</u>	<u>Achievements</u>	71
<u>10.2</u>	<u>Results</u>	72
<u>10.3</u>	<u>Discussion and recommendations</u>	73
<u>11</u>	<u>Summary</u>	77
<u>12</u>	<u>Acknowledgements</u>	78
<u>13</u>	<u>References</u>	78
<u>13.1</u>	<u>Project stage reports</u>	78
<u>13.2</u>	<u>Remaining references</u>	78

1 Introduction to the project

This is the final scientific report to be delivered from the Storm Scale Modelling Project. The project began in autumn 2002 and finished in autumn 2004. The objective was to investigate the ability of a storm scale configuration (with a 1-km grid spacing) of the Met Office NWP model to predict flood-producing rainfall up to 12 hours ahead and to develop appropriate tools for interpreting and presenting the predictions so that they have the capacity to enhance operational flood prediction capabilities.

To put this in context, the highest resolution that has been run operationally to date by the Met Office is the 12 km gridlength used in the UK mesoscale model. This is insufficient to resolve the majority of convective storms over the UK (or anywhere else). A model with a grid length of 1 km should be capable of representing many more of these storms.

The project was split into the five stages. These are briefly outlined below. The results were documented, as the project progressed, in a series of end of stage reports. The direction of the project followed a natural progression. At the start, the main concern was to examine whether a storm-scale configuration was able to produce sensible and realistic looking forecasts. Once it became established that the model does indeed have that capability, the characteristics of the model could be examined more closely. Work could then proceed on testing the sensitivity of the model to key parameter changes, generating specialised output products, developing a new approach for evaluating performance and investigating the impact of data assimilation.

1.1 The five stages

Stage 1 Initial case studies

Four case studies of different types of thunderstorm events were chosen to provide a variety of meteorological situations for testing the model. High resolution simulations (1 or 2 km gridlength nested inside a 4 km gridlength model) were run to see how realistic the forecasts were and to get a subjective assessment of how well the model performed in comparison to the operational 12 km model. In these tests the high-resolution forecasts all started from the same initial conditions as the 12 km forecasts.

Stage 2 Sensitivity studies

The sensitivity to various parameter changes was examined following issues raised during stage 1. An important development that came out of this testing was a modification to the way the model uses the convection scheme to represent convective clouds that can not be resolved on the grid. A baseline model configuration was established on the basis of results from this project and sensitivity studies performed within the High Resolution Trial Model (HRTM) project (Lean 2003).

Stage 3 Products

The primary purpose of a storm-scale model is to improve operational flood prediction capability. Post-processing of the model output will be essential if this is to be achieved successfully. Products have been developed that (1) take into account the inherent difficulty of predicting the small scales such as individual showers that a 1-km model is capable of resolving and (2) are specifically designed for use in flood-prediction.

Stage 4 Objective verification methods

There is a need to be able to assess the performance of precipitation forecasts from a high-resolution NWP model in a way that measures the accuracy of a forecast over different spatial scales. A storm-scale model is not likely to be very skilful at the grid scale, but could nevertheless provide very useful information over say an area the size of a river catchment. New methods for objectively verifying high-resolution model precipitation forecasts over different spatial scales have been developed. They involve the use of the diagnostic products presented in stage 3.

Stage 5

The verification methods developed in stage 4 were used to assess whether high-resolution (1 and 4 km gridlength) precipitation forecasts were improved by using data assimilation methods to incorporate observational information into the high-resolution model at the start. The high-resolution forecasts were also objectively compared with 12 km gridlength forecasts.

2 Why is a storm-scale model needed

Before setting out to show what a storm-scale model is capable of, we should begin by establishing why we think there is a need for such a model, both in terms of what it is we want to predict and why current forecast systems can let us down.

Convective storms are often very difficult to predict, even a few hours ahead, and can have a serious impact on society. The impact can range from poor driving conditions on roads to localised flooding if drains are unable to cope and much more serious flash flooding of larger areas if rivers burst their banks. Such floods can result in millions of pounds worth of damage and even human fatalities (e.g. the cost of the Boscastle flood has been estimated at more than £500million). The financial impact of flooding continues to increase as more and more homes and businesses are located in vulnerable locations on flood plains and the increase in household and business technology is leading to larger insurance claims.

Flash floods produced by convective storms can arrive so suddenly that it is often impossible to take action because there is very little warning. Sometimes the only indication comes from a rainfall radar picture a few minutes before the event occurs. This is the situation we need to improve. Even an extended warning of just an hour or two or a more reliable indication of the risk of a serious event occurring would be extremely valuable. It is hoped that a storm-scale model can deliver this capability. The rest of this section will involve a brief outline of the current state of the art in forecasting heavy rain and then lead on to the anticipated benefits of improving the resolution of NWP forecast models.

2.1 The current state of the art in forecasting convective storms

The operational Met Office tools available for short-range precipitation forecasting are :-

1. Nimrod and Gandolf. Advection nowcasting systems.
2. NWP forecast models with current operational horizontal gridlengths of ~60, 20 and 12 km.
3. The Convection Diagnosis Project (CDP). A model of convective showers based on post-processing of NWP 12 km fields.
4. Forecaster skill and understanding.

2.1.1 Current advection nowcasting systems

Advection nowcasting tools such as Nimrod (5 km gridlength) (Golding 1998) and Gandolf (2 km gridlength) use rainfall observed by radar as the starting point and have algorithms to advect the rainfall forward in time. After a few hours the advection forecast is gradually blended with a 12 km NWP forecast. These forecasts can be very useful up to 2-3 hours ahead, but are unfortunately least successful in the sort of situations we are interested in – i.e. when convective storms may develop. The problem is that these techniques do not incorporate an adequate way of decaying or generating new precipitation cells – they can only be moved. The most important (and most difficult) part of forecasting thunderstorms is to predict when and where they will first initiate.

Recently, a new technique has been developed within the Gandolf system to produce an ensemble of advection forecasts in which small-scale features are replaced by random noise as the forecasts progress (the STEPS system) (Pierce C et al 2004). This is designed to take into account the uncertainty in predicting the movement of areas of precipitation and give a probabilistic prediction. It is a major step forward and should be a very useful tool. However, it still suffers from the problem of not being able to initiate new showers.

2.1.2 Current operational NWP models

Current operational NWP models have the distinct advantage over advection nowcasting systems of being able to simulate the physics and dynamics of the atmosphere and are therefore able to initiate and decay areas of precipitation. That is the upside, there are however three drawbacks to using the current operational NWP system. Firstly, the location of the precipitation at the start of a forecast period can not be represented as accurately in an NWP forecast as it can in Nimrod or Gandolf because of poorer resolution and the difficulty of incorporating new observations into a complex model. Secondly, the forecasts are not available until after the start of the period of interest because of the time taken to obtain and process new observations. Both these factors mean that the first 2-3 hours are usually predicted more accurately by an advection system, and is the primary reason why nowcasting tools are used. Thirdly, an NWP model with a grid spacing of 12 km or longer is not able to resolve the vast majority of convective storms properly. A sub-grid model called a convection parametrization scheme (Met Office scheme developed by Gregory and Rowntree 1990) has to be used to represent the average effect of any convection over each grid square. The problem with convection parametrization schemes is that they are not able to model the life cycle of individual showers or the dynamical and microphysical processes that are required for the formation of secondary convective cells and the organisation of thunderstorms.

2.1.3 The Convection Diagnosis Project (CDP)

The CDP (Hand 2002) is a system that uses data from the 12-km grid-length mesoscale model, along with information about land-surface features the model can not resolve and a simple cloud life-cycle model, to indicate the likelihood of showers of particular intensities occurring. It can, on some occasions, improve a forecast of convective storms considerably. In particular, it can provide useful information about the probability of convection when conditions for its initiation are marginal. However, it is still limited to the information provided by the 12 km model and tends to be most useful when showers are initiated because of land surface effects (e.g. the effect of a hill or coastline) without having the ability to respond adequately to some other dynamical causes.

2.1.4 Forecaster skill and interpretation

Human forecasters have the potential to improve a forecast by adding their own experience and meteorological understanding to the systems described above. They have the most recent observational information (satellite, radar) to hand and are able to interpret and process diverse information quickly. Nevertheless, they are still very reliant on output from the NWP forecasts and nowcasting systems, and to depart too much from those predictions is risky. The development of more accurate NWP or nowcasting systems with appropriate output products is thought to be the best route towards improved human predictions.

2.2 The anticipated benefit of a storm-scale model

In terms of predicting convective rainfall events, a storm-scale model has nearly all the advantages of a coarser-resolution NWP model, without the problems arising from insufficient resolution. The anticipated benefit comes primarily from a better representation of the structure and dynamics of showers or thunderstorms and from a higher-resolution representation of land-surface fields. There are four key reasons why we should expect a storm-scale model to give better results.

(1) The major advantage a 1-km gridlength model has over the current operational 12km model is that a convection parametrization scheme is probably no longer required (or at least far less important and restricted to shallow convection only). Unlike the 12-km model, a 1-km model should be capable of propagating showers inland during the winter, as well as (more particularly in the summer) initiating new storm cells and organising convections into larger thunderstorms, squall lines or comma clouds.

(2) The model is capable of a much more accurate representation of surface terrain. Mountains, hills, coastlines, forests, urban areas, lakes can all be much better represented. Storms are often thought to be tied to urban heat islands, sea-breeze convergence fronts, ascent or heating associated with hills or other local effects. It is thought that if these effects can be more accurately simulated, a larger fraction of convective storms can be more accurately predicted.

(3) A grid spacing of ~1km will allow the simulation of rainfall rates that are directly comparable with those measured by radar and rain gauges. The benefit of that is that it should be possible to make direct use of model output to warn of the possibility of local flash floods and the hazards associated with intense rainfall rates and then verify those forecasts.

(4) A denser grid spacing means that it is worthwhile to include and develop more accurate representation of cloud structure and microphysics. A better representation of clouds and precipitation should give us more realistic simulations of convective storms. E.g. the formation of gust fronts from evaporative cooling that can lead to the initiation of new cells.

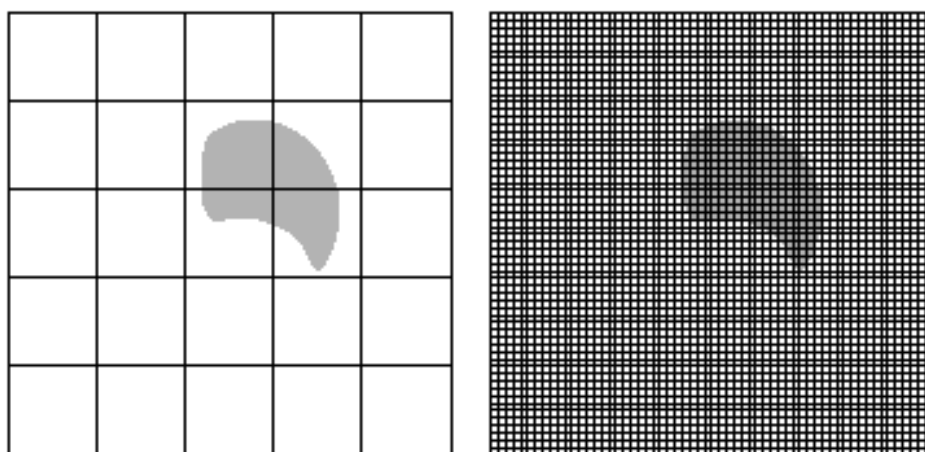


Figure 1. A schematic representation of a 12-km grid spacing (left) and a 1-km grid spacing (right). The shaded area represents a shower.

Figure 1 shows the immediate advantage of changing from a grid spacing of 12 km to 1 km. The schematic shower (~15km across) can only be represented on one or two grid-

points by the 12-km grid. That is nowhere near enough to capture the structure of the shower. But on the 1-km grid, the shower is represented by more than 150 points.

It was stated earlier that a storm-scale model has nearly all of the advantages of a coarser-resolution model, which implies that there are some disadvantages or difficulties that it would be unfair not to mention. These issues are listed below and will be raised again in later sections.

- (1) Computational expense. A higher resolution model is more expensive to run and since we already have the problem of having to wait for model output, this may be a serious consideration.
- (2) Domain size. The fact that a high-resolution model domain has to be nested inside a coarser resolution domain, along with the additional computational expense, means that that storm-scale model domain may have to be considerably smaller than the current operational 12-km gridlength domain. It also means a potentially difficult choice of where the domain should be. If a domain is too small, the high-resolution model forecasts may end up being too largely determined by the flow through its boundaries from a coarser-resolution (say 12 km) model.
- (3) Predictability. We will be attempting to forecast smaller scales and these scales are inherently less predictable. It means that post-processing will certainly be required to make sense of the output.
- (4) Science. New scientific challenges need to be addressed. The process of assimilating observations into the model at higher resolution and the representation of turbulent mixing in convective clouds when the convection scheme is switched off are examples of the problems that need to be high on the research agenda.

Despite the difficulties, it is thought that the benefits of running a storm-scale model should considerably outweigh the drawbacks. Advances in research and computer power will ultimately reduce the significance of issues (1) (2) and (4) outlined above.

3 The set up of a storm-scale model

3.1 Future ambition and current testing

Recent and anticipated increases in computer power at the Met Office should mean that an operational storm-scale (~1 km grid-length) model is a viable proposition by the end of the decade. It is expected that computing resources will be sufficient to allow a domain of a few hundred by a few hundred kilometres in size to be practical by that time. In the meantime, a project is underway to construct a 4-km grid-length model, which will cover the UK and become operational during 2005/6. It will be capable of providing the boundary conditions necessary for a storm-scale model.

At present we do not have the resources to run a 1-km model in real time with an adequately sized domain. However, we are able to run a test configuration within a research framework to investigate how such a model behaves. Initial ideas about the way this test model should be set up were based on only a limited experience of running at such high resolution. Since then, sensitivity experiments have allowed us to make more informed decisions about the best choices of model parameters and formulation, and shed light on important issues or failings. This has enabled us to arrive at a baseline configuration for a high-resolution modelling system, which is being used as a standard from which further developments can take place. The ultimate aim is that, several iterations down the line, the baseline configuration will become a new operational storm-scale model.

3.2 Collaboration with the High Resolution Trial Model (HRTM) project

At this point, the role of the HRTM project must be acknowledged. The HRTM project was set up to construct a high-resolution modelling system for testing in trials (Lean 2003). A large proportion of the sensitivity studies have been done within the HRTM framework. Collaboration with the HRTM project has made the storm-scale modelling project much more successful and enabled progress to be made in several areas in both projects. For example, it would have been much more difficult to assess the capability of a storm-scale model without having access to a reference model and trial data. In turn, the HRTM project has benefited from the model assessment, parameter testing, modifications to the convection scheme and new precipitation verification techniques that have arisen from the storm-scale modelling project.

3.3 The model configuration – Technical decisions

The decisions that were (and continue to be) made about the set up of the model can be broadly split into two categories.

1. Technical
2. Scientific

Having said that, scientific reasoning should always go hand in hand with technical decision making and technical considerations will emerge alongside new scientific developments, but it is a useful split for the purposes of this report.

Decisions had to be made at the start about the most appropriate size and location of domains and the most sensible horizontal grid-spacing to use. The values of other important parameters were arrived at by means of systematic sensitivity studies. The default values that were eventually settled upon for a base-line configuration are given in Table 1. They are regarded as suitable for now, although there are likely to be further modifications in future as the system evolves and if more appropriate alternatives emerge.

The sub-sections below describe how the values in Table 1 were arrived at for the current system and what the expectations are for future changes.

	Horizontal Grid length		
	12 km	4 km	1 km
Domain size (km)	~ 1740 x 2172	~ 756 x 756	~ 300 x 300
Number of grid points	26572	36100	90000
Number of vertical levels	38	38	76
Time step	5 minutes	1 minute	30 seconds
Boundary updating every (minutes)	60	30	15
Diffusion	None	Del-4, e-folding time 8 time steps Coeff 1.14e4	Del-4, e-folding time 8 timesteps Coeff 1.43e3
Critical Relative humidity (RH-crit)	85%	85%	85%

Table 1. Parameter settings arrived at as the defaults following sensitivity studies.

3.3.1 Domain size, domain location and horizontal resolution

The current HTRM domains are shown in Figure 2. These are the domains that have been used for most of the case studies in this project, but some of the earlier simulations were performed on different domains (see the stage-1 report).

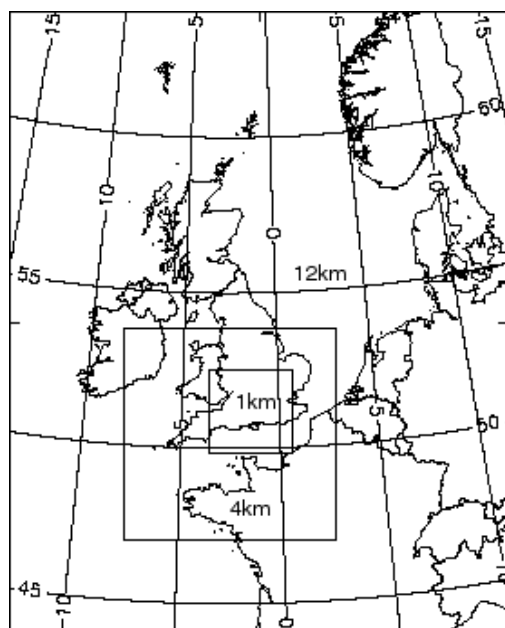


Figure 2. The 1 and 4-km domains used for most of the case studies. The 12-km domain is the same as that used operationally.

The 1-km gridlength domain was chosen because it is the largest we can test with a reasonable turn-around time, given current computer resources. It is also similar to the size of domain we might hope to use in the first operational system. The horizontal grid spacing of 1 km is necessary because it is the coarsest resolution we can reasonably refer to as ‘storm-scale’. The 4-km gridlength domain was primarily designed to ‘house’ the 1-km domain, and for that reason, and computational speed, it was made as small as we could get away with. Any operational 4-km domain would need to be substantially larger to cover most or all of the UK.

3.3.2 Vertical resolution

Current configuration

Two sets of vertical levels were tested for both the 4 and 1-km models. Firstly, the set of 38 levels used operationally in the ~60-km global model and 12-km mesoscale model was adopted, then a doubling to 76 levels (an extra level between each existing level) was examined. The levels are not equally spaced; both sets have many more levels low down (a few tens of metres spacing near the ground) than towards the top (a few kilometres spacing in the stratosphere). Results from sensitivity studies of convective events, using the two sets of levels, were largely inconclusive. Mostly, there was little difference in forecasts, but on some occasions there was a noticeable impact. The choice of the number of vertical levels is also complicated by other factors and it still needs ongoing investigation. For the purposes of this work it was decided to use 38 levels as the default in the 4-km model so that data assimilation could more easily be applied and because it is cheaper to run. However, for the 1-km model it is regarded as physically more appropriate to use 76 levels especially if slantwise structures such as fronts are present (Lean and Clark 2003, Persson and Warner 1995).

Future plans

Further examination of the behaviour of the 1-km model, in particular, with different vertical resolutions is necessary. This will require detailed investigation of why differences arise and a close look at the impact on convective development. Ultimately, we want to use more than 38 levels, so it is essential to find out how this can best be done and what other areas of the model formulation have an impact on the decision. It is likely that at some stage the operational 12-km model will have more vertical levels, and this will feed into the storm-scale model development.

3.3.3 Time step

Current settings

The time step is the time interval over which the equations in the dynamics part of the model are solved. A longer time step is more economical but less accurate, so the choice here involves dealing with the trade-off between accuracy and efficiency. The higher the resolution the shorter time step must be to maintain sufficient accuracy. It is partly why a high resolution model is more expensive to run. The choices for the 4 and 1-km grid-length models given in Table 1 are considered to be the best compromise between accuracy and expense for test simulations run in a research context.

Future plans

It is possible that the time steps could be lengthened to reduce costs in an operational system. This requires further testing to make sure that the results are not compromised too much and model is robust enough to cope.

3.3.4 Diffusion

Current values

The diffusion equation can be used to apply smoothing to temperature, humidity and wind fields. However, the use of diffusion is sometimes contentious. The operational 12-km model is run without any diffusion because it is thought that it should not be needed with a semi-lagrangian formulation for the dynamics and we do not wish to remove important detail from the forecasts. At higher resolution the addition of a small amount of horizontal diffusion shown in table 1 has been necessary to reduce the occurrence of unrealistic grid-scale precipitation patterns associated with convection. The difficulty is that when the convection scheme is switched off, as it is for a 1-km grid length forecast, any showers simulated by the model dynamics can either have a tendency to develop into unphysical single grid-point storms if there is no diffusion or be delayed in initiating for too long if diffusion is added. The amount of diffusion currently added represents a compromise between the two opposite behaviours.

Future plans

What is really needed is a representation of the small scale mixing associated with convection (in the free troposphere above 1 km or so) on the grid-squares where convection occurs with little or no diffusion applied elsewhere. This will have the desired effect of removing unrealistic ‘grid-point storms’ whilst not delaying convective initiation. One way of doing this would be to switch the convection scheme back on, but this is a backward step in a high-resolution model, unless the scheme can be modified in a suitable way. A modified convection scheme has been tried for the 4-km simulations and will be discussed later. Another alternative is to use the ‘targeted diffusion’ that has very recently been introduced into the 12-km model. It works by adding diffusion where vertical velocities are highest, i.e. where there are updrafts into convective clouds. The third, and perhaps best approach, is to implement a more physically realistic representation of turbulence where convection occurs. This is a current area of research and development at JCMm.

3.3.5 Boundary Updating

Current approach

The operational 12-km grid-length model supplies the values that are needed at the boundaries of the 4-km model, which in turn supplies the boundary information for the 1-km model. The outer few points of a domain (the rim) are fixed to values from the coarser outer model. Immediately inside that rim, the information supplied from the coarser model is reduced linearly over a few points until the contribution is zero.

Ideally, the boundaries should be updated as often as possible (every time step) to eliminate significant mismatches between what the inner and outer models think is happening at the inner-model boundary and inhibit the development of any spurious features that may result. The boundary updating intervals shown in Table 1 have been

chosen because they are the longest that give satisfactory results in most situations. The constraint has been the demand on computer storage that frequent updating demands.

Future plans

In future, more frequent boundary updating should be possible and may be required. Another ongoing project at JCMM, which would circumvent the issue of boundary updating, is to develop a variable resolution grid. This would involve the use of a single grid with the highest resolution (~1 km grid spacing) in the main region of interest and this would gradually degrade to coarser resolution elsewhere.

3.3.6 The critical relative humidity (RH-crit)

Current settings

This is the relative humidity at which the model diagnoses that some cloud has formed in a grid box (Smith 1990) and is less than 100%. The reason for doing this, (rather than diagnose clouds when the relative humidity is very close to 100%), is to take account of variability within grid squares. If the average relative humidity is say 90% over a grid square, it is likely that the value will be close to 100% in some places within that square. It is reasonable to suppose that values of RH-crit should become closer to 100% as the size of the grid squares is reduced, because there can be less variability within smaller squares, but it is difficult to know what those values should be. There was concern that forecasts might be very sensitive to changes in this parameter so experiments were performed to investigate the impact. It turned out that changes to RH-crit made some differences to the detail, but not the essence of the forecasts in the convective cases examined. For that reason, the values of 85% at most model levels used in the 12-km mesoscale model were retained for the higher resolution grids.

Future testing

It is possible that this will need to be re-assessed, especially in the light of other changes to the representation of clouds and precipitation in the model.

3.4 The model configuration – Scientific decisions

3.4.1 The representation of convection

Current methodology

This sub-section gives an overview of the work that was presented in the stage 1 and stage 2 reports. The stage 2 report, in particular, discusses the use of a modification to the convection scheme.

The vast majority of convective storms can not be resolved properly by the operational 12-km grid-length mesoscale model. This has meant that a convection parametrization scheme (Gregory & Rowntree 1990) has been required to represent the average effect of showers within each grid box. Without it, unrealistic storms can develop at single grid points which may ruin a forecast or even cause the model to fail. The convection scheme is used to firstly determine whether convection will initiate, and then to calculate the required adjustments to the temperature and moisture profile that will follow any

convection. Unfortunately, because a convection scheme represents a snapshot of convective clouds in equilibrium with their environment, it is incapable of modelling the development and decay of showers or the dynamics that leads to convective organisation. It can only (if it is working correctly) produce a rainfall picture that is a smoothed average over an area, rather than develop individual showers. The realism and usefulness of forecasts can suffer as a result.

The big advantage of going to a 1-km grid-length model is that most convective showers can then be resolved by the model dynamics. It should be possible for the model to simulate the initiation, structure and decay of individual convective events. However, it is still not possible at that resolution to resolve the smaller showers or storm cells and it is conceivable that there will be a particular difficulty in simulating the very first stages of shower development. The question that then arises is whether there is still a need to use a convection scheme in a 1-km grid-length model? The current view is that it is better to switch it off, as we want the model dynamics to do the work unhindered. Certainly we should not use the convection scheme as it stands because it is not designed for such high resolution.

Results from case studies (section 4) have shown that a 1-km grid-length model with no convection scheme (but with numerical diffusion) does produce realistic forecasts. For that reason, it is currently the default to switch off the convection scheme for 1-km simulations.

The situation becomes very different when the grid spacing is 4-km. At that resolution, larger convective systems will be resolved by the model, but most convective events can not be resolved properly, particularly at the early stages of development. It would seem likely that a convection scheme is still required to prevent the aliasing of too much energy into single grid point storms. However, the convection scheme is not designed for that resolution and there are concerns about how it will interact with the model dynamics when convection is partly resolved.

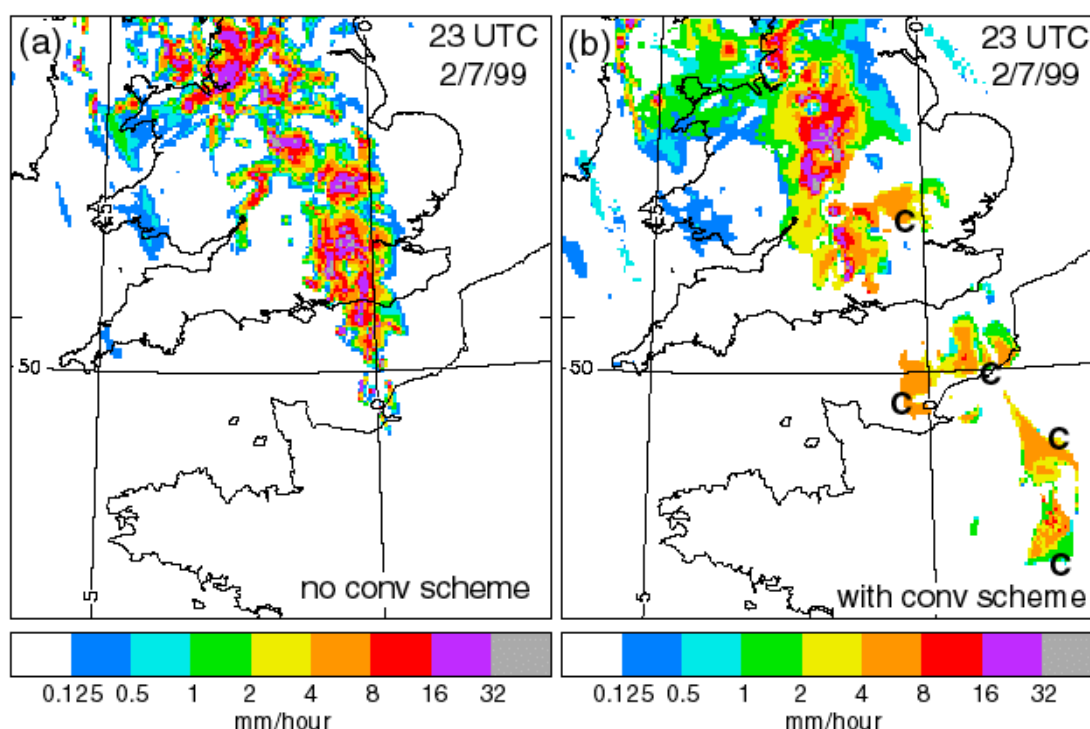


Figure 3. Precipitation rates at 23 UTC on 2nd July 1999 from 8-hour forecasts run with a grid spacing of 4 km. (a) convection scheme switched off. (b) convection scheme included. Rain areas labelled 'C' come from the convection scheme.

Case studies have been run to investigate the behaviour of a 4-km grid-length model when run with and without the convection scheme. The results have confirmed that neither gives satisfactory results and another solution is needed. Figure 3 shows an example of the very large differences that can be found between a forecast with the convection scheme included and a forecast without. In this particular example of an intense convective event the convection scheme triggered too much and in the wrong place, as shown by the rain areas labelled 'C'. The forecast without the convection scheme was much better. Figure 4 shows another example from a less intense convective situation when scattered showers developed and became more organised during the afternoon and evening. In this case the forecast with the convection scheme switched off produced far too much rain in the afternoon after initiating too late. The forecast with the convection scheme switched on started off well, but then the convection was not maintained into the afternoon when the heavier organised showers occurred.

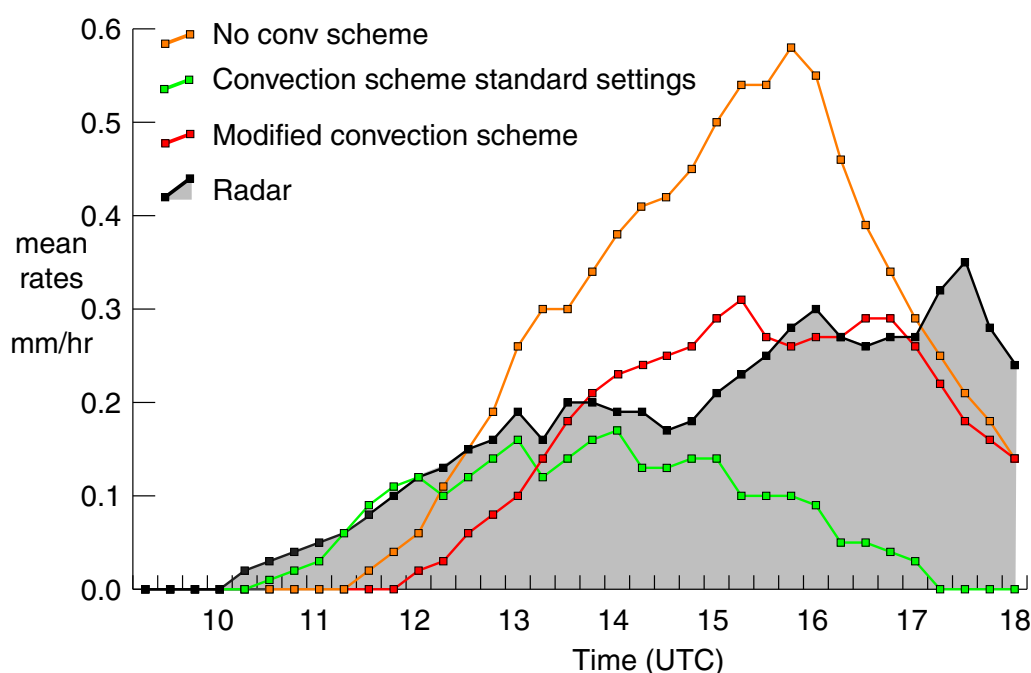


Figure 4 . Graph of the mean rainfall rates from radar (black line and shading) and 4-km gridlength forecasts (coloured lines) over the 1-km domain shown in Figure 2, for 3rd May 2002. The forecasts started from 01 UTC and were run with either the convection scheme switched on, switched off or modified as indicated by the colour key.

The problem with the convection scheme was that it was too active for a grid spacing of 4 km. It tended to remove the convective instability before the dynamics had a chance to develop any showers. A modification to the convection scheme has been made so that a limit is put how active it is allowed to be. This was presented in the stage 2 report. It has improved test simulations and is now used in the default 4-km configuration. Figure 4 shows how the rainfall rates become more realistic in the afternoon (red line) when the modification was applied because convective organisation was allowed to take place. This is a positive result because it is the larger heavier showers we are most interested in from the point of view of flood prediction. The other positive aspect is that the same modification when applied to the forecast shown in Figure 3 inhibited the development of spurious rain from the convection scheme and produced a forecast similar to that with no convection scheme. The downside is that the delay in initiation became worse in the 3rd May case (Figure 4).

It might seem strange to have concentrated so much on modifying the convection scheme in the 4-km model when it is the 1-km model we are primarily interested in, but there are good reasons for doing so. Firstly, the 4-km model supplies the information at the boundaries of the 1-km domain and if the wrong information is coming through the boundaries the 1-km forecast will give poor forecasts. Secondly, the approach can also be applied to the 1-km model (or any model with a grid-length greater than ~0.5 km and less than ~12 km). Thirdly, we should bear in mind that a ~4-km model will become operational first.

Future plans

The problem of representing convection correctly in a high-resolution model still requires considerable research and investigation. The modification to the convection scheme at 4-km has helped and the 1-km simulations have produced some impressive results without a convection scheme, but there are still underlying issues to be addressed.

The basic issue is how to represent only those scales that can not be properly resolved and leave all the rest to the dynamics. Up till now the options available have been to use a convection scheme or add numerical diffusion. The problem with a convection scheme is that it is not designed for use on a high-resolution grid even when modified to make it more suitable. Also, a convection scheme will act before the convection can be simulated on the grid and therefore tends to inhibit the resolved showers from triggering. Diffusion does the same thing; by acting everywhere even before convection has broken out, it delays the resolved triggering.

A new alternative could be to apply a 'targeted diffusion' which acts only where the vertical velocity in resolved convective cells exceeds a particular threshold. This should have the beneficial effect of removing undesirable single grid-point storms without delaying the onset in of resolved convection, but does not deal with the initiation stage or with turbulence in weaker showers (and puts an arbitrary 'switch' into the system). As previously mentioned in the sub-section about turbulence, research is now underway to develop a more scientifically valid way of representing the turbulence associated with convection. It may then be possible to replace the need for a convection scheme at both 4 and 1 km with this new approach.

3.4.2 Data assimilation - Getting the best possible analysis

Current approach

The 'analysis' is the name given to the start of a forecast. It should follow that a more accurate analysis will lead to a more accurate forecast. For that reason, we must aim to give each forecast the best possible start. In the context of a storm-scale modelling system, this is both very important and technically difficult to do. It is important because the primary aim of a storm-scale model is to predict localised rainfall events and it is unlikely to be successful in doing that if the model is incapable of getting the rainfall structure correct to start with. The particular difficulty we face lies in the fact that we are trying to represent such small scales in a storm-scale model. It is not good enough to get the rainfall broadly correct; details in the precipitation pattern such as individual storms do matter and need to be captured. Not only that, a rainfall analysis must be consistent with the model dynamics, otherwise the information will not be retained into a forecast.

The start of a forecast is adjusted to be as close as possible to available observations by means of a process called data assimilation. There are a number of data assimilation methods currently available and the development of new approaches is an active area of research. The techniques applied in the operational 12-km grid-length UK mesoscale model have been described in the stage 4 report. They are:-

1. 3DVAR. A technique for adjusting model winds, temperatures and humidities to be closer to observed values. (Lorenz et al 2000)
2. MOPS – Latent heat nudging. A method for altering rainfall pattern in the model by adjusting temperatures according to where rain is observed by radar. (Jones and Macpherson 1997)
3. MOPS – Cloud analysis. This is a technique for adjusting the cloud in a model to give a better fit to the observed cloud structure using observations from satellite, radar and surface observations (Macpherson 1996).

At the start of the storm-scale modelling project data assimilation was not available to use in either the 4 or 1-km gridlength models. The first set of case-study experiments, documented in the stage 1 report, used the 12-km model analysis interpolated to the high-resolution grids as the starting point for high-resolution forecasts. This was appropriate at the time because it was useful to examine the impact of extra resolution on a forecast when the initial conditions were the same. However, it was also unsatisfactory because the 4 and 1-km gridlength models had to ‘spin up’ detail over the first few hours of every forecast.

Data assimilation was included at 4 km for testing during winter of 2003/2004. To start with, this meant implementing the same methods used at 12 km in the 4-km model. The techniques were only modified to the extent they needed to be to make them work. It was recognised that further developments were necessary, but the first objective was to examine how well the data assimilation performed when first introduced. The same methods were not possible at 1 km. As an alternative first step, the updated information obtained from running 3DVAR at 4 km was incorporated into the 1-km model. The data assimilation methods that were used are presented in Table 2.

	First set of case studies Used for stages 1 and 2	Second set of case studies Used for stage 5 Current default setup
12 km	3DVAR MOPS latent heat nudging MOPS cloud analysis	3DVAR MOPS latent heat nudging MOPS cloud analysis
4 km	No 4-km data assimilation Interpolation from 12 km	3DVAR MOPS latent heat nudging MOPS cloud analysis
1 km	No 1-km data assimilation Interpolation from 12 km	No 1-km data assimilation Addition of 4-km 3DVAR increments (MOPS introduced since stage 5)

Table 2. The data assimilation methods used in the case studies.

The future

Current avenues of research at JCMM for high-resolution data assimilation are as follows

1. Further work on the implementation of 3DVAR at 4 km.
2. Introduction of MOPS latent heat nudging and cloud analysis in the 1-km gridlength model.
3. The use of Doppler winds from radar in 3DVAR.
4. New methods for using radar by relating rainfall pattern to model dynamics.
5. Research into the use of infrared satellite imagery in 3DVAR
6. The use of shorter assimilation cycles – i.e. more frequent data assimilation and forecasts
7. The use of more advanced assimilation techniques such as 4DVAR and ensemble Kalman filters

3.4.3 The representation of clouds and precipitation

Current approach

The way that clouds and precipitation are represented by the 12-km grid-length mesoscale model has also been used in the 4 and 1-km model simulations. The cloud microphysics is modelled by a so called 'mixed phase bulk water scheme' (Wilson and Ballard 1999), which represents cloud water droplets, rain, ice/snow particles and the microphysical interactions between each particle type. The only modification to the representation of clouds that was tested during the experiments was the impact of changing the value of RH-crit, which determines the cloud fraction within a grid square. The effect of changing RH-crit has already been discussed.

New developments

The representation of cloud and precipitation processes within the 12-km grid-length model is a very much simplified version of what we think happens in reality. It is has to be that way for three reasons:-

1. To limit the computational expense so that the model does not take too long to run.
2. Because the size of each grid square is too coarse to warrant more complexity.
3. Because the addition of more complexity is not necessarily straightforward, especially given our incomplete understanding of real clouds.

When the resolution is increased to a 1-km grid-spacing, we have to reconsider whether the second point in the list is still valid. Two changes have occurred that may cast some doubt on that premise. Firstly, individual convective cells can now be resolved, and secondly, rain can be blown across more than one grid square. Now that showers can be resolved, it may become important to represent graupel (soft hail) as this is frequently generated within convective clouds. Work is in progress at JCMM to introduce graupel into the scheme as an additional cloud-ice variable. Work has also been carried out to specifically distinguish between small pristine ice crystals and larger aggregates of crystals (snowflakes).

The other main issue that needs consideration is the way rain is represented. The current operational approach is to diagnose, at each time step, when rain is occurring in each grid square, by using an assumption that the rainfall remains constant over a large number of time-steps. Such an approach is not suitable when clouds can be resolved and are rapidly evolving. It means that convective rain will fall directly through strong updrafts instead of being carried upwards, and this may act to weaken showers before they can organise properly. It also means that rain may fall into the wrong grid squares if it is not allowed to be blown from one to another. The addition of a capability to allow rain to be blown by the wind has been developed and has produced encouraging results. Figure 5 is a picture produced by Richard Forbes at JCMM which shows an example of a particular type of situation in which the more realistic treatment of rain has improved the distribution of rain falling into individual river catchments. It is now available for use and further testing. References are Forbes and Halliwell 2003 & Wilson and Forbes 2004.

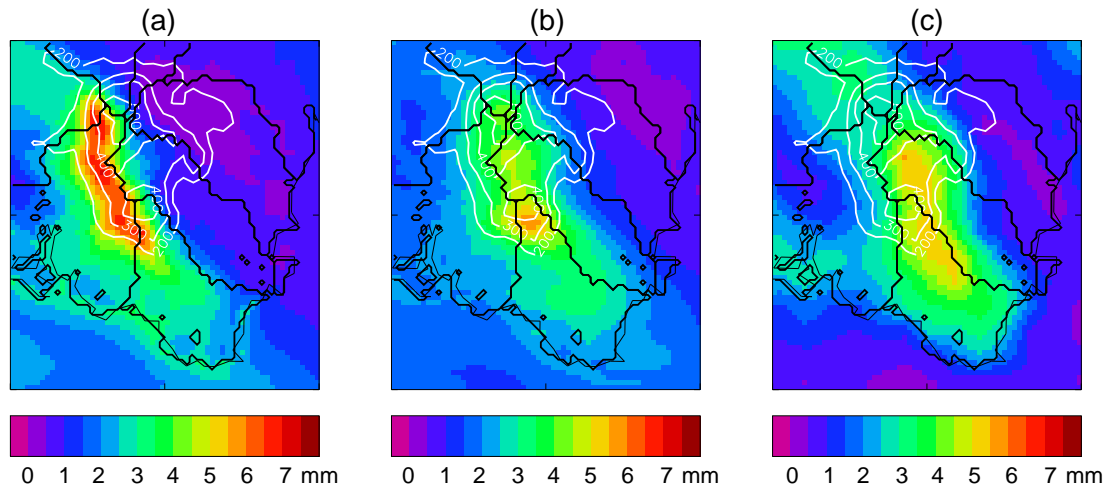


Figure 5. Illustration of the impact of including a prognostic rain variable on the orographic enhancement of rain in a 2 km resolution version of the model. The figure shows 3-hour normalised rainfall accumulations (in mm) on 29 Nov 2001 and the river catchments associated with Dartmoor in south-west England (orography shown in white, river catchment boundaries and coastline shown in black). (a) Model forecast with diagnostic rain, (b) model forecast with prognostic rain and (c) radar estimates of rainfall accumulations. The orographic enhancement is largely through the feeder/seeder process; the downwind drift of rain is sufficient to significantly change the catchment into which the rain falls. Courtesy of Richard Forbes.

Of course, extra complexity also means extra cost. The task that lies ahead is to find the most appropriate way of representing cloud and precipitation in a storm-scale model without making the model too expensive to run as a forecasting tool. Other factors that also need serious consideration are how any changes to the cloud microphysics will interact with any modifications to other areas of the model such as the convection parametrization or a new turbulence scheme. This will become more of an issue when the new prognostic cloud fraction scheme, which has recently been introduced into the global-area model, is tested in the storm-scale model. The new scheme gives a better representation of sub-grid cloud so that processes such as convection and turbulence can have a direct impact on the cloud fraction.

3.5 The model configuration - representation of the land surface

Current configuration

A high-resolution grid allows a much more detailed representation of orography (and coastlines). Figure 6 shows the vast improvement in the representation of hills and valleys in a 1-km grid-length model compared to that used at 12 km. It should be an important factor in determining how well a model is capable of representing local weather events that are linked to variations in the height of the terrain. The orography used comes from a 1-km resolution Digital Terrain Elevation Data (DTED) dataset (developed by the National Imagery and Mapping Agency (NIMA, formerly Defence Mapping Agency (DMA))). It has been smoothed to an effective resolution of more like 2 km. The reason for doing this is to prevent the generation of spurious grid-scale features in a forecast. Sensitivity studies of the impact of orographic smoothing have not been performed in this project, but might provide a useful insight into the triggering mechanism of some storms.

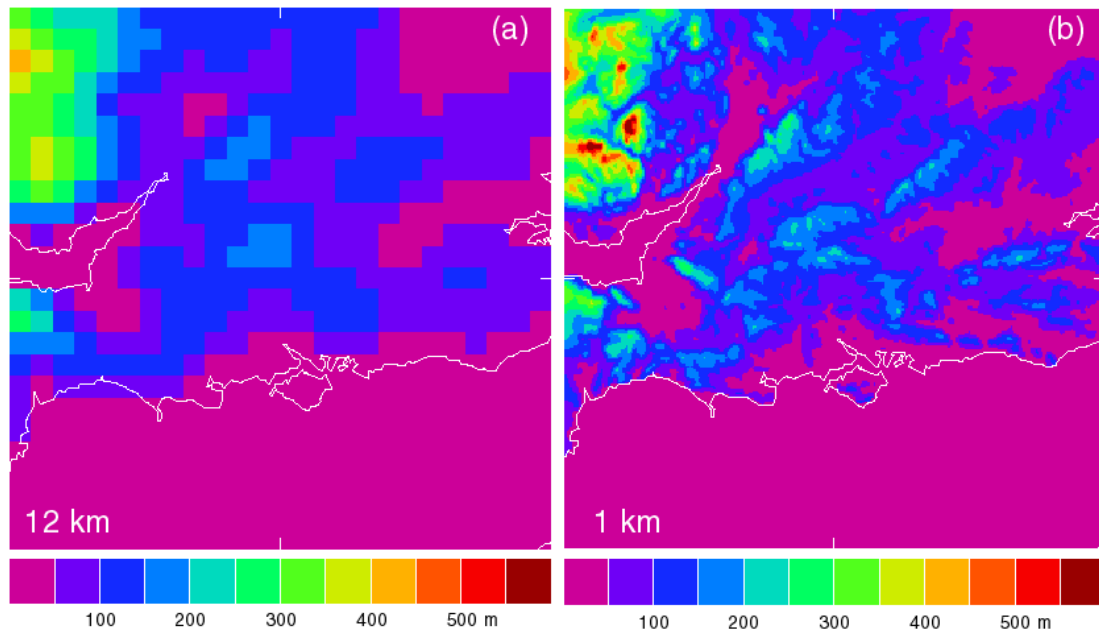


Figure 6. The orography used in (a) a 12-km gridlength model and (b) a 1-km gridlength model.

The representation of what is on the land surface is taken from a 100 m resolution dataset produced by the Institute of Terrestrial Ecology (ITE). It is used to obtain fractions of particular land-types within each 1x1 or 4x4 km grid square. Nine land types are defined; broad leaf trees, needle leaf trees, short grass, tall grass, shrubs, urban, water, soil and ice, which are used within the Met Office Surface Exchange Scheme (MOSES) (Cox et al 2001).

At present, the amount of moisture in the soil is the same as that used in the 12-km operational model. It is updated weekly.

Future plans

The main advance needed is an improvement to the accuracy of the soil-moisture. A way to do this is to make use of the recent development of a surface and subsurface hydrology diagnosis facility within the Nimrod nowcasting system. This uses a modified version of MOSES along with the Probability Distributed Moisture (PDM) scheme developed at the Centre for Ecology and Hydrology (CEH) (CEH Wallingford 2001). It is updated using hourly rainfall accumulations from quality controlled radar data and currently provides soil moisture fields on a grid spacing of 5x5 km. The impact on high-resolution-model rainfall forecasts of including a more accurate representation of soil moisture on high-resolution-model rainfall forecasts is yet to be seen. It should have a positive effect, but this needs to be determined through sensitivity studies.

4 Encouraging results from initial case studies

The first objective of the storm-scale modelling project was to run high-resolution simulations of a few interesting events and examine in detail how well the model performed. Four cases were run and the results documented in the stage 1 report. When the cases were run, the storm-scale model was at a very early stage of development. No additional information was included through data assimilation at the start of the high-resolution simulations. The aim was to see whether there was evidence to suggest that such a model has the potential to deliver significantly better rainfall forecasts than is currently possible, even though not fully developed at the time of testing.

4.1 The Case studies

1. 2-3rd July 1999. Rapidly moving severe thunderstorms and squall line with some places receiving >50mm of rain in 1 hour.
2. 11-12th October 2000. A quasi-stationary band of convective rain over Sussex and Kent that produced severe flooding.
3. 3rd May 2002. Scattered convection and thunderstorms with some organisation.
4. 29th July 2002. Isolated thunderstorms over East Anglia with flash floods.

The reasons for choosing these particular case studies were because they represented different weather situations, they all produced heavy rain, and they were associated with varying success in operational forecasts.

Case 1 was well forecast by the operational 12-km model, cases 2 and 3 were moderately well forecast and case 4 was poorly forecast. Their diversity allowed us to examine how the model was able to represent different types of convective situations. Three of the cases produced flash flooding. Table 3 is included to show the diversity of the four events.

A summary of the findings from each of the cases is now presented. The stage 1 report gives a much more comprehensive description.

	Case 1	Case 2	Case 3	Case 4
Scattered convection			Yes	
Isolated convection				Yes
Embedded convection		Yes		
Organised structures	Yes		Yes	Yes
Squall line	Yes			
Mesoscale convective system	Yes			Maybe
Mostly dynamically driven	Yes	Yes		Yes
Mostly heating driven			Yes	
Severe	Yes			Yes
Moderate		Yes	Yes	
Lightning	Yes	Yes	Yes	Yes
Flooding reported	Yes	Yes		Yes

Table 3. Characteristics of each of the four initial case studies.

4.2 Case 1, 2-3 July 1999

Figure 7 shows the radar picture of the storm at 00 and 03 UTC. Model simulations of the event were run using horizontal grid-lengths of 12, 4 and 2 km.

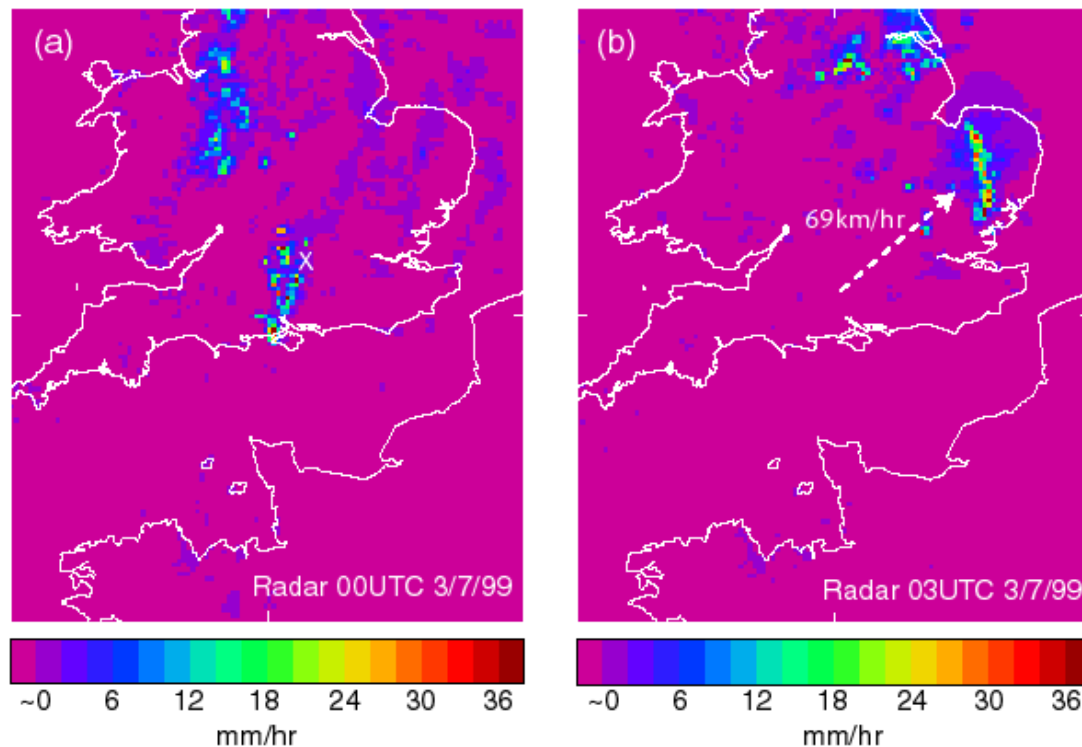


Figure 7. Rainfall rates from radar at 00 and 03 UTC 3rd July 1999. 'X' marks the location of Reading where the gauge at the University site recorded more than 40 mm of rain fell in less than 40 minutes.

The operational 12 km forecast.

The operational 12-km forecast was regarded as being good as it gave a strong signal for organised convective precipitation. It had good timing of the event and got the spatial extent of the storms reasonably well. This was a difficult forecast to improve upon with higher resolution.

The 4 and 2 km forecasts

Improvements over the 12 km forecast

1. Both the 4 and 2-km forecasts produced a squall line structure that looked very much like that seen in radar imagery. Figure 8(b) shows the 2-km simulation. The 12-km forecast was unable to produce this kind of structure as realistically (even though it appears to have done in the snapshot shown in Figure 8).
2. They both propagated the rainfall at a speed much closer to the 69km/hour observed. The 2-km forecast was the closest.
3. They both produced more accurate precipitation rates and accumulations. Again, the 2-km forecast was the better of the two.

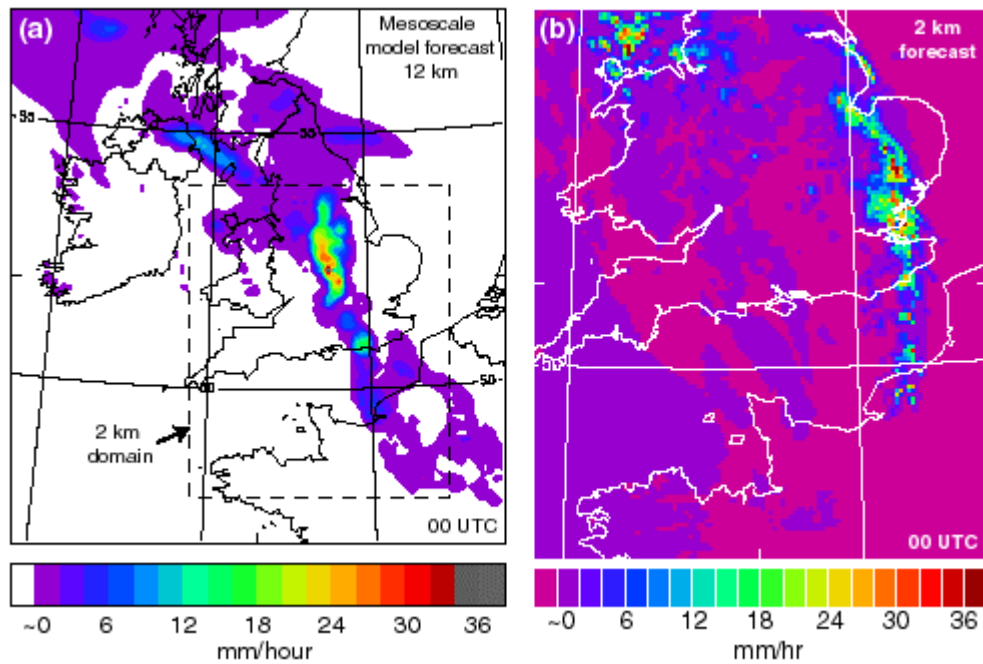


Figure 8. Rainfall rates at 00 UTC 3rd July 1999 from (a) mesoscale model (12-km gridlength) forecast and (b) 2-km gridlength forecast. Both forecasts began at 9 hours earlier.

Issues

There was a significant timing error. Although the high-resolution forecasts had a better propagation speed for the storms, they arrived 2-3 hours too early. Figure 8 shows that the storms predicted by the 2-km forecast at 00UTC are located where they were observed by radar at 03UTC (Figure 7). In that sense the higher resolution simulations were less accurate than the operational forecast. However this is misleading as the 12-km forecast was better for the wrong reasons. A closer inspection revealed that the 12-km forecast appeared to be better because it has compensated for the incorrect timing of an upper-level frontal zone by generating the storms in the wrong place. The upper-level front was incorrectly timed at all resolutions, so the high-resolution forecasts were more dynamically consistent.

Another concern was the sensitivity of the 4 and 2-km grid-length forecasts to whether the convection scheme was switched on or not. The forecasts with the convection scheme were unable to develop a squall line structure and produced other spurious bands of rain. Even though it is open to debate whether a convection scheme should be employed at these resolutions, it is clear that better results were obtained when it was switched off in this particular example.

Overall impression

The higher resolution forecasts gave better results in terms of the structure of the precipitation, the propagation speed and the rainfall intensity than the operational 12-km model. The spatial accuracy was comparable. The timing was worse, but this was determined by the larger-scale flow through the domain boundaries and the 12-km was only better because of compensating errors. The 2-km grid-length model with the convection scheme switched off predicted the most realistic rainfall rates and totals.

4.3 Case 2, 11-12th October 2000

This was one of the famous flood events of autumn 2000. A quasi-stationary band of heavy showers and thunderstorms persisted for more than 15 hours and produced local rainfall accumulations in excess of 100 mm. The model was run with a gridlength of 12, 4 and 2km.

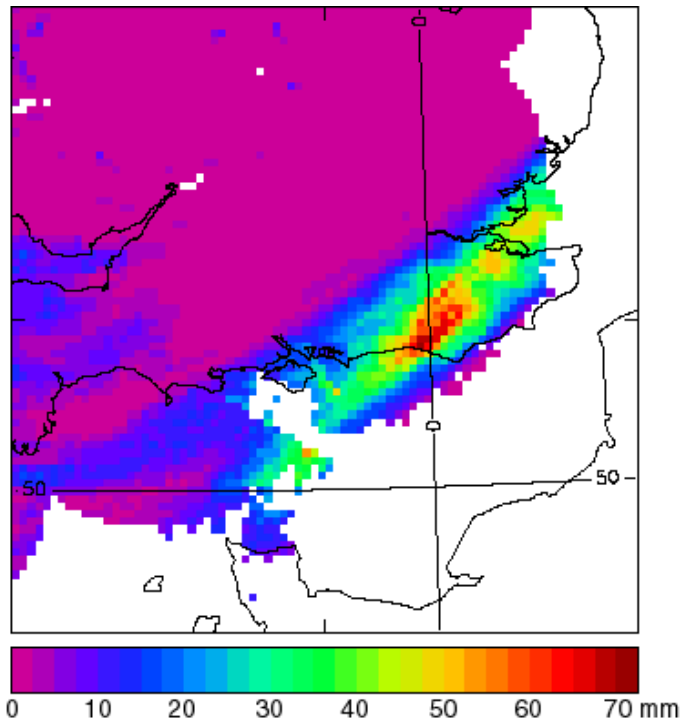


Figure 9. Rainfall accumulations from radar for the 15-hour period from 19 UTC 11th October 2000 to 10 UTC 12th October 2000.

The operational 12-km forecast

The operational forecast did produce a quasi-stationary band of showery precipitation, but the rainfall accumulations were substantially lower than those observed and the band was located some 30-50km to the northwest of where it should have been. The forecast did not provide a warning that flood-producing rainfall totals were a risk over Essex and Kent.

The 4 and 2-km forecasts

Improvements over the 12 km forecast

The 4-km forecasts did produce some improvement. The rainfall accumulations were slightly larger and the location of the band was slightly further southeast.

The 2-km forecast was considerably better. The rainfall accumulations over 12x12km squares increased from ~35mm in the 12-km forecast to ~60mm in the 2-km runs. Not only that, the band of high accumulations was 20-30 km further southeast and much of it now coincided with the observed region of high accumulations. Much of the improvement can be put down to the 2-km model being able to produce a much more accurate representation of the storm cells that occurred within the band. The 12-km and to some extent the 4-km model produced most of their rain from the convection scheme,

which is not designed to simulate the individual showers and tended to trigger too much along the south coast.

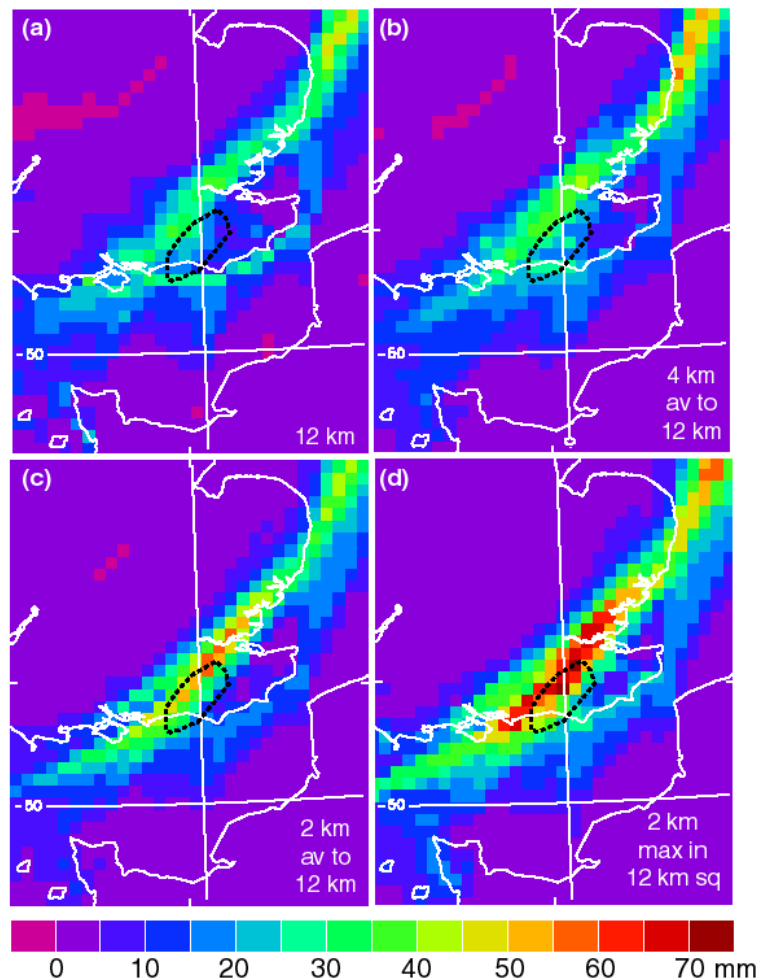


Figure 10. Rainfall accumulations over the period 18 UTC 11th to 12 UTC 12th October 2000. (a) 12-km gridlength forecast. (b) & (c) 4 and 2-km forecasts averaged to the 12-km grid. (d) maximum accumulation values from the 2-km forecast within 12x12km squares. The forecasts all began at 15 UTC 11th October. The black line encloses the area where accumulations from the radar exceeded ~55mm.

Issues

The sensitivity of the forecasts to whether the convection scheme was used remained an issue. The pictures in Figure 10 are all from forecasts that included the convection scheme. Simulations were also run with the convection scheme switched off and with the new modification to the convection scheme. These forecasts did vary from the runs with the convection scheme, but the signal for a dramatic improvement in skill with the 2-km model remained.

Overall impression

The 2-km gridlength forecasts (with and without the convection scheme) were significantly better than the 12-km operational forecast. Rainfall accumulations were much closer to that observed and the high accumulations were more accurately positioned. The 4-km forecasts gave less of an improvement.

4.4 Case 3, 3rd May 2002

This was the least dramatic event of the four, though still with heavy rain. Scattered showers developed widely and became heavy in places before becoming organised into a band of thunderstorms that tracked across the London area.

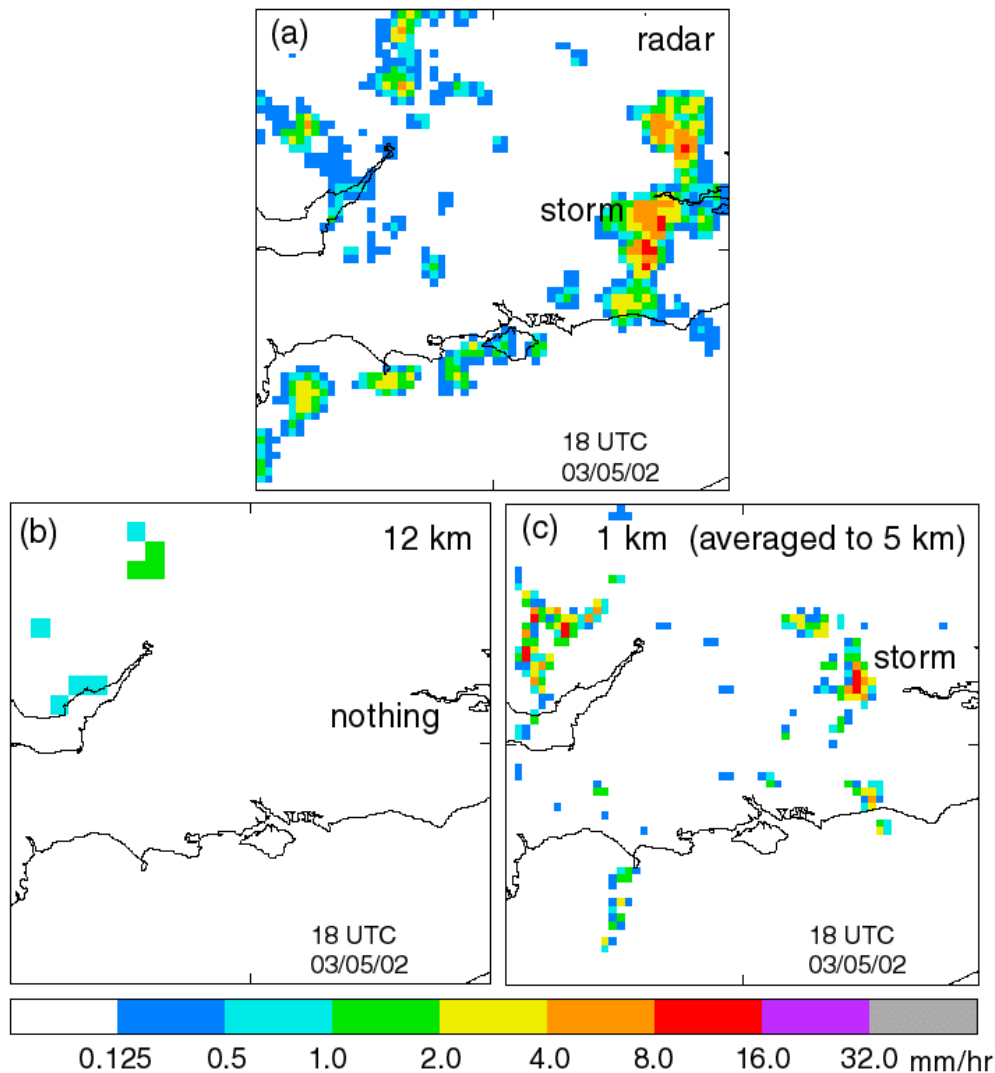


Figure 11. Rainfall rates at 18 UTC 3rd May 2002 from (a) radar (b) & (c) 18-hour forecasts from 12 and 1-km gridlength models.

The operational forecast performed well in some respects but not in others. On the plus side, it did produce scattered showers over England during the morning and into the afternoon. However, it was incapable of predicting the observed rainfall rates because of the limitations of resolution and more importantly, it completely failed to develop the band of thunderstorms over southeast England in the evening. Figure 11 shows the deficiency of the 12-km forecast at 18 UTC.

The 1 and 4-km forecasts

Improvements over the 12 km forecast

The 1-km forecast was distinctly better than the 12-km forecast in two respects.

1. It produced rainfall rates that were similar to those observed
2. It maintained organised larger showers into the evening. Figure 11 shows the showers in the 1-km forecast at 18 UTC. They were not as widespread as observed, but it is nevertheless a big improvement on the 12-km forecast.

The 4-km forecast that included the new modification to restrict the convections scheme (not shown) also retained organised showers into the evening, but with poorer spatial accuracy than the 1-km forecast.

Issues

Two new issues came to light from the high-resolution forecasts of this event.

1. There was a delay in triggering the showers in the morning. They initiated around 30 to 45 minutes too late in the 1-km forecast and 60 to 90 minutes too late in the 4-km forecast. The delay is a result of instability having to build up over a whole grid square before a shower can form, and is therefore worse at coarser resolution in any simulations with the convection scheme switched off or restricted. Such a delay in triggering does not occur at 12-km because the showers are then represented by a convection scheme. However, the 12-km forecast was actually no better in timing shower initiation because the convection was triggered more than an hour too early instead.
2. Showers were absent from areas close to the edge of the domain where flow was coming into the domain through that boundary. Figure 11 shows that there was very little precipitation in the 1-km forecast at 18UTC within the northern 30-40km of the domain. This was the case throughout the forecast period. The problem is that the showers needed more time to develop (or 'spin up') in the flow that had recently advected in from a coarser-resolution domain. A similar effect could be seen at the edge of the 4-km domain, but the problem was much less obvious because the domain was so much larger.

Overall impression

This case study has highlighted one of the main benefits of a 'storm-scale' resolution model. The 1-km grid-length forecast was able to resolve individual showers and therefore simulate the transition from small scattered showers into more organised thunderstorms. The 12-km operational model was unable to do this, and as a result, produced a poor forecast of evening downpours.

The inhibition of shower formation near the northern boundary of the 1-km model is of concern and needs to be examined further.

The 4-km gridlength model could also simulate the transition from scattered to organised convection as long as a suitable restriction to the convection scheme was applied. However, it was not as realistic or spatially accurate as the 1-km model and the initial delay in shower formation was worse.

4.5 Case 4, 29th July 2002

An isolated but severe thunderstorm complex developed to the east of London during the night and tracked north across East Anglia in the morning. Figure 12 shows snapshots of the rainfall from radar at 05 and 13 UTC. Flash flooding was reported.

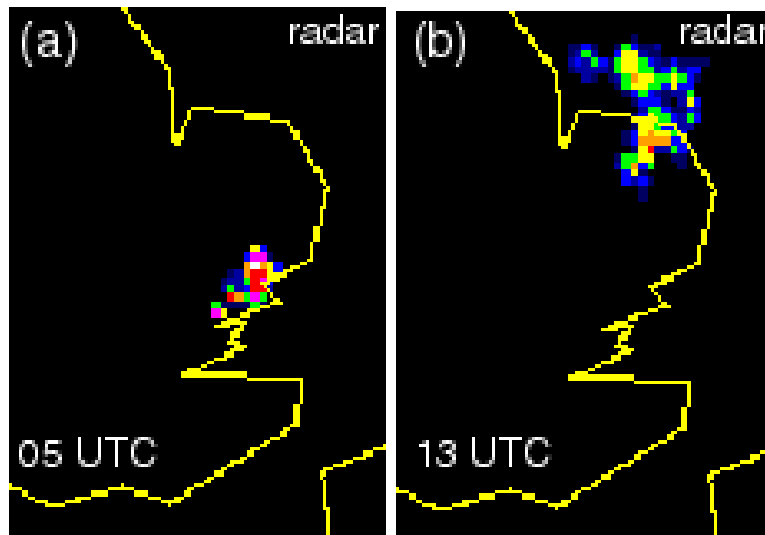


Figure 12. Rainfall rates from radar at 05 and 13 UTC 29th July 2002. Rainfall rates are in mm /hour and the colours are blues 0.125 to 1, green 1 to 2, yellow 2 to 4, orange 4 to 8, red 8 to 16 mm, pink 16 to 32 and white >32mm/hour.

The 12-km forecast

The 12-km grid-length model produced a poor forecast. It was unable to generate any thunderstorms and therefore gave no indication whatsoever that such an intense storm might develop (see Figure 13(a)).

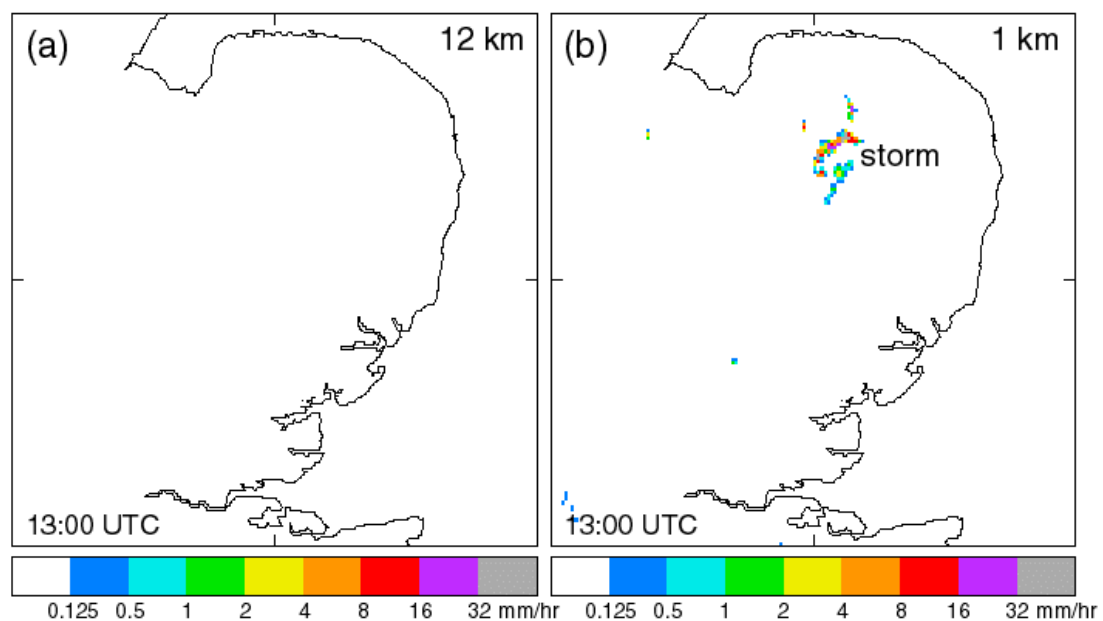


Figure 13. Rainfall rates at 13 UTC 29th July 2002 from (a) a 12-km gridlength forecast and (b) a 1.2-km gridlength forecast. Both forecast started at 00 UTC 29th July.

The 1 and 4-km forecasts

Improvements over the 12 km forecast

The 4-km forecast was no better than the 12-km forecast. It also failed to generate a storm.

Figure 13(b) shows that the 1-km forecast was able to produce a thunderstorm over the correct area. It had a realistic-looking structure, spatial extent and rainfall rates. However, it did develop the storm several hours too late.

Issues

An issue to highlight from this case was the sensitivity of the 1-km grid-length forecast to the number of vertical levels that were used. With 38 levels (the same as that used operationally at 12 km), the storm developed as shown in Figure 13(b), but with 76 levels (double the 38 levels) no coherent storm developed. The reason is not known, but it is possible to speculate that the use of 76 levels enabled a stronger capping inversion to form and that this provided more inhibition to convective triggering.

Overall impression

The 1-km grid-length model (with 38 levels) provided the only forecast that was able to predict this event. Such a result adds weight to the expectation that a storm-scale model does have the potential to improve our ability to predict severe storms that can lead to unexpected flash flooding.

The case has also shown that further investigation is needed into the sensitivity of high-resolution precipitation forecasts to changes in vertical resolution.

4.6 Summary of the results from the case studies

The main message to be taken away from the subjective assessment of the first four case studies is that the performance of the high-resolution (1 or 2 km grid length) simulations was very encouraging. In three of the cases, the highest resolution forecasts were considerably better than the 12-km forecast model in predicting the location, structure and intensity of convective rainfall events. Even for case 1 when the 12-km model performed well, the higher resolution was able to produce a more realistic rainfall structure and intensity.

The intermediate resolution forecasts (4-km grid spacing) also performed somewhat better than the 12-km model. However, there was some concern about the difficulty such a model has in resolving smaller showers and in the effect the convection scheme has at that resolution. A modification to the way the convection scheme is used has been implemented and has had a positive impact, although the fundamental difficulty of using a resolution that only partially resolves many convective storms remains a problem.

Other issues that the case studies highlighted as requiring further attention are:-

- (1) The sensitivity to vertical resolution. We need to know whether storm-scale model performance is generally sensitive to the choice of vertical levels, why the sensitivity arises and whether there is any systematic behaviour.
- (2) The 'spin up' problem close to the edge of the domain.
- (3) The delay in the initial triggering of convection.

4.6.1 Subjective performance scores

In an attempt to make the assessment more objective in nature Table 4 has been constructed to provide scores for fundamental characteristics of the forecasts. These have then been combined to give an average score for each of the model resolutions. The 1 and 2-km forecasts have been combined in this process because otherwise the figures would be unbalanced for such a small sample of diverse events.

The individual scores are from 0 to 5.

0 = no skill
1 = very poor
2 = poor
3 = OK
4 = good
5 = extremely good

The four categories are:

1. Rainfall accumulations – how well did the forecast produce the observed accumulations somewhere in the area of interest.
2. Spatial accuracy – did the forecast produce the rain in the correct place.
3. Temporal accuracy – did the forecast produce the rain at the correct time.
4. Precipitation structure – how well did the forecast simulate the correct precipitation structures (e.g. squall line, scattered showers, comma cloud, embedded frontal convection etc).

	Model grid spacing															
	12km				4 km				2 km				1 km			
	Cases				Cases				Cases				Cases			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Rainfall accumulations	3	1	2	0	4	2	3	0	4	5	-	-	-	-	4	2
Spatial accuracy	3	1	3	0	3	2	4	0	3	4	-	-	-	-	3	3
Temporal accuracy	5	5	1	0	3	5	3	0	3	5	-	-	-	-	4	2
Precipitation structure	3	3	1	0	4	4	3	0	4	4	-	-	-	-	4	5
Average scores	1.94				2.50				3.69							

Table 4. Subjective scores for various aspects of forecast performance for the resolutions examined in the four case studies.

The average scores shown in Table 4 reinforce the perception that the 1 and 2-km grid-length forecasts of convective rainfall events were significantly better on average over the four cases than the 12-km forecasts. These results are particularly noteworthy when it is remembered that the high-resolution model used was still at an early stage of development and did not have the benefit of extra data assimilation at high-resolution.

However, a note of caution is also needed, the conclusions are being drawn from only four case studies. To assess the model performance properly we also require objective verification scores from a large number of cases. Preliminary results from an objective verification method are presented in section 8.

5 Products from a storm-scale model

The case studies have indicated that a storm-scale model does indeed have the potential to deliver more accurate forecasts of high-impact rainfall events than our current operational systems can achieve. Attention must now turn to how we can make best use of the storm-scale model output for heavy rain and flood prediction. To make best use of this advance in forecasting capability, there is a need to develop appropriate products that are designed to meet customers' requirements and expectations and also address the particular problems that arise in the interpretation of high-resolution model output.

An explanation of why we need to post-process the output from a storm-scale model is given below. After that, some examples of the kind of products that have been produced are shown, with a focus on the requirements for flood prediction. This is a shortened version of the much more comprehensive material presented in the stage 3 interim report.

5.1 Why are post-processed products required

The reason why the post-processing of high-resolution precipitation forecasts is necessary can be covered under three headings.

1. Forecast uncertainty. Numerical model forecasts are always associated with some uncertainty.
2. Unpredictable scales. A high resolution model will attempt to predict some scales that may be inherently unpredictable for a given length of forecast.
3. Customer requirements. Customers will benefit greatly from products that are tailored to their needs (and higher resolution allows more scope for a greater variety of useful products).

1. Forecast uncertainty

It is not possible to represent the exact state of the atmosphere at the start of a forecast (whatever resolution). Any initial fields can only be a best estimate with an associated error. If forecasts are run from almost identical initial states, they will diverge and the differences can be explained in terms of a forecast 'error' or 'uncertainty'. Since the initial state (including its error) is never completely known, we cannot regard any single forecast as the most likely possibility; it is only one realisation from an infinite number of alternatives with different likelihoods that form a probability distribution. In a nested high-resolution modelling system, further uncertainty can be introduced because coarser-resolution information is passed through the boundaries into the higher-resolution grid.

But why does this mean that post-processed products are required? The answer is a matter of honesty. If we know that our forecast system has some degree of uncertainty associated with it, then we should be presenting that uncertainty in some way.

2. Unpredictable scales

One of the major concerns often raised about a storm-scale model is that it will be attempting to resolve features that are inherently unpredictable within the time period over which forecasts will be run. In other words, a model with a grid spacing of 1 km will have the capacity to resolve quite small showers, but the exact location of any individual shower is not predictable after a short period of time (perhaps less than an hour sometimes). This element of forecast uncertainty can also be a factor in coarser-resolution

forecasts, but is a particular problem in high-resolution modelling. The problem is that for the smallest most unpredictable scales, forecast errors can grow so fast that eventually the error on those scales becomes the same for all forecasts regardless of the initial state, and can be regarded as random noise. It means that we should not believe the fine-scale detail in a high-resolution precipitation forecast. In making this statement, we are not saying that a high-resolution forecast system can have no skill, rather, that we should be concentrating on the most predictable aspects such as rainfall accumulations over larger areas and the characteristics of the precipitation features.

In considering how to post-process high-resolution precipitation forecasts we should consider how to deal with two scales. (1) The scale over which we expect well-resolved features to be in error. (2) The scale over which we need to average to account for small-scale noise and partly resolved features. We want to find the smallest scale that is generally predictable for a particular forecast system. Then we can generate products that will provide as much detail as we think the model is capable of predicting to an acceptable level of accuracy. The difficulty is that this scale may vary greatly from event to event.

3. Customer/user requirements

It is very unlikely that the raw precipitation output from a high-resolution NWP forecast model will be the most useful way that users/customers can have access to the information. It makes sense to provide something more suited to customers' individual requirements. For example, the Environment Agency will be concerned about the possibility of flash flooding and could use products that focus on rainfall amounts over susceptible river catchments. A local authority or emergency service might want specific information about likely peak rainfall intensities on a stretch of motorway and not be so interested in the amount of rain. Organisers of a sporting event might want to know how likely it is to rain at that particular location that day and for how long if it starts.

5.2 Examples of forecast products for flood prediction

A selection of forecast products will now be presented.

The 11-12th October 2000 flood event (case 2 in section 4) is used as a means of showing some of the kinds of forecast products that could be produced from a high-resolution forecast system for use in flood prediction. The pictures have been generated from output from 2-km grid-length simulations of the event. A full description of how each specific product was generated is given in the stage 3 report. This set of examples is not exhaustive, it is intended to give an impression some of the possibilities and highlight products that are already available for incorporation into a storm-scale modelling system. Although the pictures are displayed in a form designed for human interpretation, the same products could also be incorporated into an entirely automated warning system.

5.2.1 Products displayed on square areas

The first product that might be automatically generated is a picture of the average rainfall accumulations that are forecast to occur over square areas within a particular time period. Squares are used for the sake of simplicity and because we know from previous arguments that we should not present the output on the grid scale. An example of such a product has already been shown in section 4 in Figure 10 (b & c). In that example a square of size 12x12 km was chosen to give a direct comparison with 12-km grid-length output. The size of the squares should depend primarily on the spatial accuracy of the forecast system and on user

requirements. The most suitable square size for a particular forecast system has to be determined by objective forecast verification and this is covered in section 6.

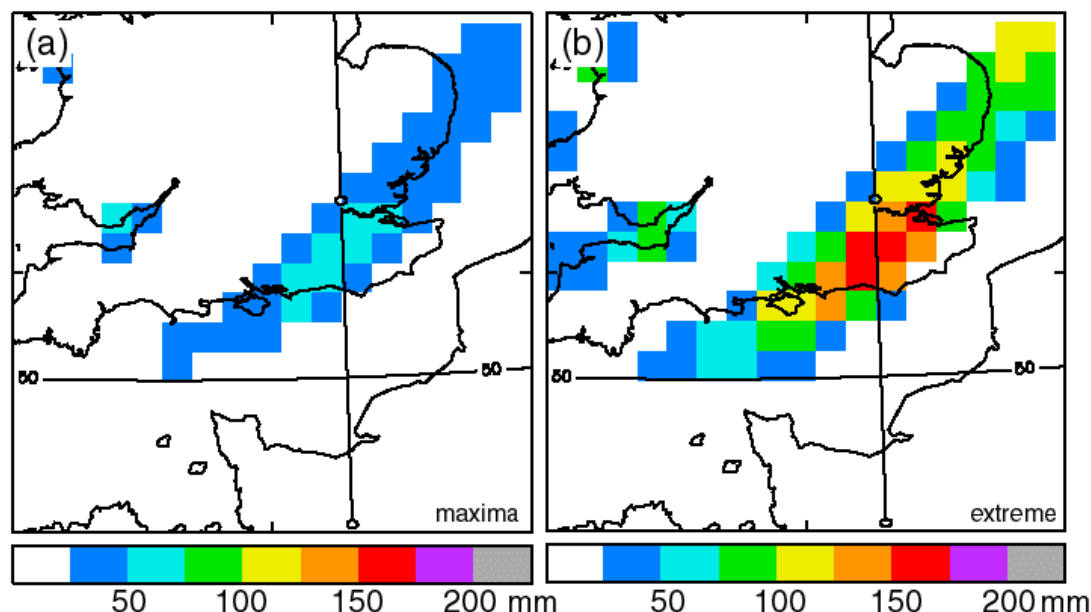


Figure 14. Products from 2-km gridlength rainfall forecasts for the 18-hour period 18 UTC 11th Oct 2000 to 12 UTC 12th Oct 2000. (a) The maximum accumulation generated within 30x30 km squares. (b) The extreme accumulation within 30x30 km squares. See text for more information.

The next piece of information that might be required is an indication of what the highest accumulations look like. In addition to showing the average rainfall totals over squares, the maxima could also be presented over the same sized squares and over the same period. Figure 10 (section 4) and Figure 14(a) give an example of this product (from two different forecasts). They clearly show that the maximum accumulations can be considerably higher than the average and are more likely to be much larger where there are extreme localised events. It is a valid way of extracting the more reliable detail from the unpredictable scales because we are sampling over space and time. It shows that there might be high localised accumulations without being too specific about where. The same approach can also be used for rainfall rates to give a picture of where the heaviest rain might occur.

It is possible to go a stage further. Since we know that any forecast has an associated error and the smallest scales have low predictability, we could find a worst-case scenario in which the heaviest rainfall within a shower (or rain area) repeatedly falls in the same place. Such situations do occur in nature when storms develop at the same convergence point over a period of time, and these events can lead to flash flooding. Figure 14(b) gives an example of such a product in which the extreme accumulations are displayed. At short intervals (every timestep or few minutes) the rainfall rates are converted into accumulations. Then a cell detection algorithm is used to separate out individual rain areas. Once the individual rain areas/showers are identified, the accumulations are adjusted to take account of the possibility that the highest accumulations within a particular shower could occur elsewhere within that shower (within a defined radius). As this is meant to represent a rare extreme, it can not easily be verified, but it is interesting to note that the extreme values presented in Figure 14(b) and for the Boscastle flood event (see later) are comparable to those that actually occurred.

For the purpose of visual interpretation, the presentation of output on squares tiled as shown in Figure 14 is fine. If, however, the product is to be part of an automated warning system then a variation using overlapping squares (see stage 3 report) is more suitable because it then matters much less where the squares are positioned.

5.2.2 Products for river catchments

To predict flash floods, we need to show how much rain is forecast to fall into individual river catchment areas. It is possible to do this with high-resolution model output (~ 4 km grid-length or less) because there will be a sufficient number of grid squares within even the smaller catchments (particularly from a 1-km model). Figure 15(a) shows a basic product of average rainfall accumulations within river catchments over a period of time.

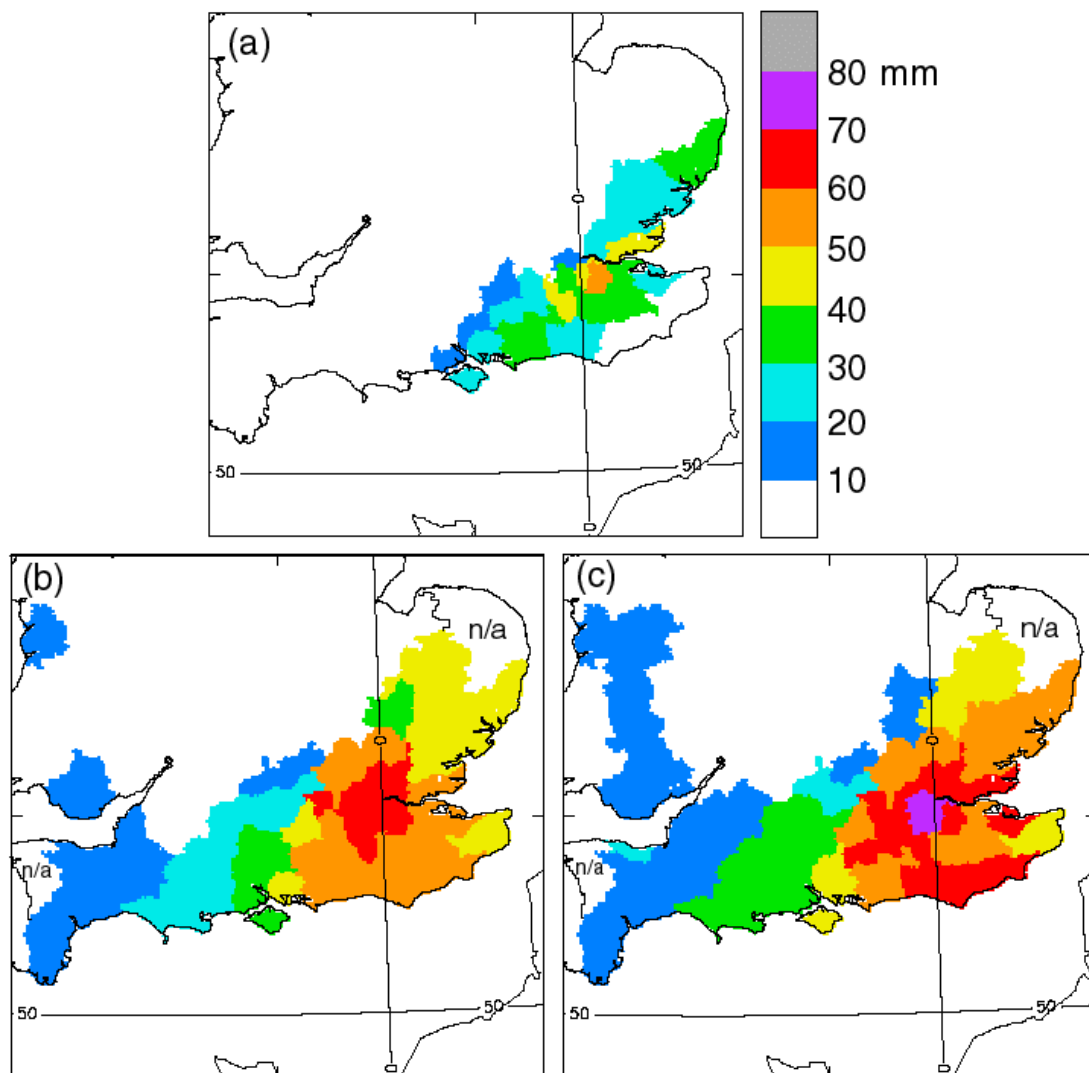


Figure 15. River catchment products from a 2-km gridlength rainfall forecast for the 18-hour period 18 UTC 11th Oct 2000 to 12 UTC 12th Oct 2000. (a) Catchment average rainfall accumulations. (b) Maximum possible catchment average accumulations given a spatial uncertainty of 40 km. (d) Maximum possible accumulations over half-catchment areas given a spatial uncertainty of 40 km. See text for more detail.

The presentation of catchment-average accumulations is useful, but is extracted from a single forecast realisation, which we know has errors. We need to take account of the possibility that different forecast scenarios may result in considerably higher accumulations over some catchments. For example, it is possible that if a predicted area of high rainfall totals straddles two catchments and the model has a displacement error, the reality could be that all the rain falls into one of the catchments (and leads to a flash flood). Figure 15(b) shows a product that displays a worst-case scenario for catchment-average rainfall accumulations after taking into account a forecast displacement error of up to 40 km. This highest average accumulation is found by sorting the rainfall accumulations within a larger region surrounding each catchment (extended some distance, 40km in this case, beyond a catchment boundary) so that the highest valued pixels within the larger area that are equal in number to the number of pixels in the catchment itself are added together and divided by the number of pixels in the catchment.

This worst-case scenario approach can be taken a stage further. Instead of computing the maximum average accumulation that could occur within an entire catchment, we can do the same but for an area that is half the size of a catchment (or some other specified sub-catchment area) but still display over the whole catchment. The reason for doing this is to pick out the potential for more localised high-accumulation rainfall events that could have an impact within a sub-catchment of a larger catchment. An example is shown in Figure 15(c). The information is still presented over the larger catchments to take account of the uncertainty.

All of the catchment-average products can be included in an automated warning system.

5.2.3 The use of probabilities

A particularly good way of presenting forecast uncertainty is to generate probabilities. Figure 16(a) shows an example of a product which displays the probability that a particular rainfall accumulation will be exceeded within a period of time.

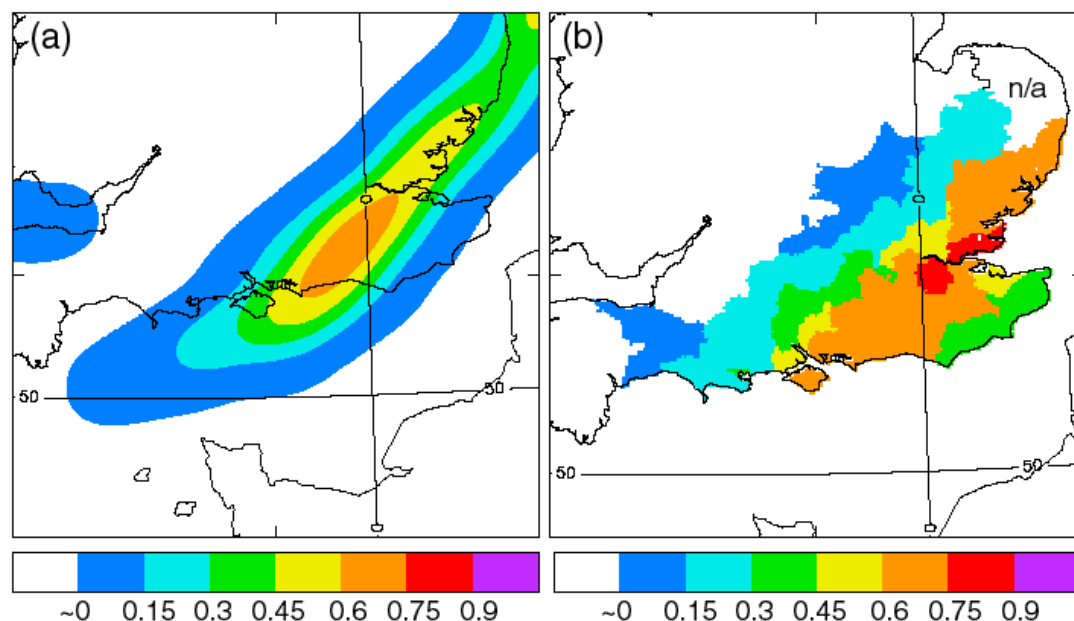


Figure 16. Probability products from a 2-km gridlength rainfall forecast for the 18-hour period 18 UTC 11th Oct 2000 to 12 UTC 12th Oct 2000. (a) The probability that the rainfall accumulation at each model grid square will exceed 30 mm. (b) The probability that the catchment-average rainfall accumulation will exceed 20 mm.

The probabilities were found at each grid-point (in this example) by first finding the fraction of points within a circle surrounding each grid-point that exceed the accumulation threshold, and then applying a recursive filter to spread out those probabilities. This particular example only took account of the spatial error. Other methods can also be used to account for spatial or intensity errors. The result should be a smooth picture of the chance of a particular amount of rain falling within a period of time.

Probabilities can also be converted into odds (or risk) to give a better depiction of a small chance. Rainfall rates can be treated in much the same way as accumulations. The use of probabilities can also be extended to river catchments. Figure 16(b) shows such a product; again it is for a particular accumulation threshold being exceeded over a time period. This time the probabilities were found by moving each catchment area, grid square by grid square, around the domain up to a specified distance from where they came from and computing the catchment-average accumulation at each location, then finding the fraction of occasions on which the threshold accumulation was exceeded.

All of the probability products can be included in an automated warning system.

5.3 Hydrological applications

The products for river catchments need not just be displayed as pictures, they can be more specifically designed for hydrological applications. Figure 17 shows how the average accumulation over a catchment can be presented as cumulative hourly totals. Two lines are drawn in this example; one is the predicted cumulative average accumulation, the other gives a worst case scenario if there were to be some realistic displacement error in the predicted rainfall.

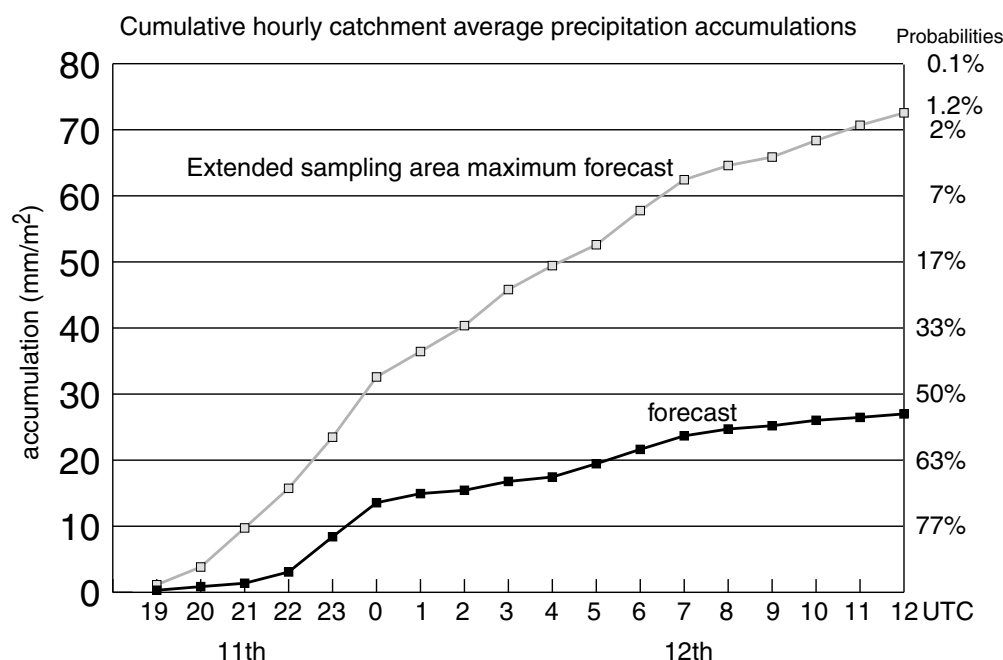


Figure 17. Graphical product from a 2-km gridlength rainfall forecast for the 18-hour period 18 UTC 11th Oct 2000 to 12 UTC 12th Oct 2000. Graph of the hourly cumulative catchment-average rainfall accumulations for a particular catchment. The black line is from the actual forecast; the grey line is a worse-case scenario. The probabilities on the right are for exceeding 18-hour accumulations thresholds.

Other lines could be added for the worst-case rainfall totals given different forecast displacement errors. The probabilities on the right have been computed by the same approach used to generate Figure 16(b). Probabilities might be more usefully presented as hourly percentile values for catchment-average accumulations in the format shown in Figure 18.

All these graphical products could be used in an automated warning system. They also have the added advantage of providing information in a way that is suitable for input into a hydrological rainfall-runoff model. The hourly catchment-average accumulations could be fed directly into such a model. The additional benefit of generating different scenarios is that they too can be fed into a rainfall runoff/river-flow model to give an indication of what the uncertainty in a forecast could mean in terms of flooding potential.

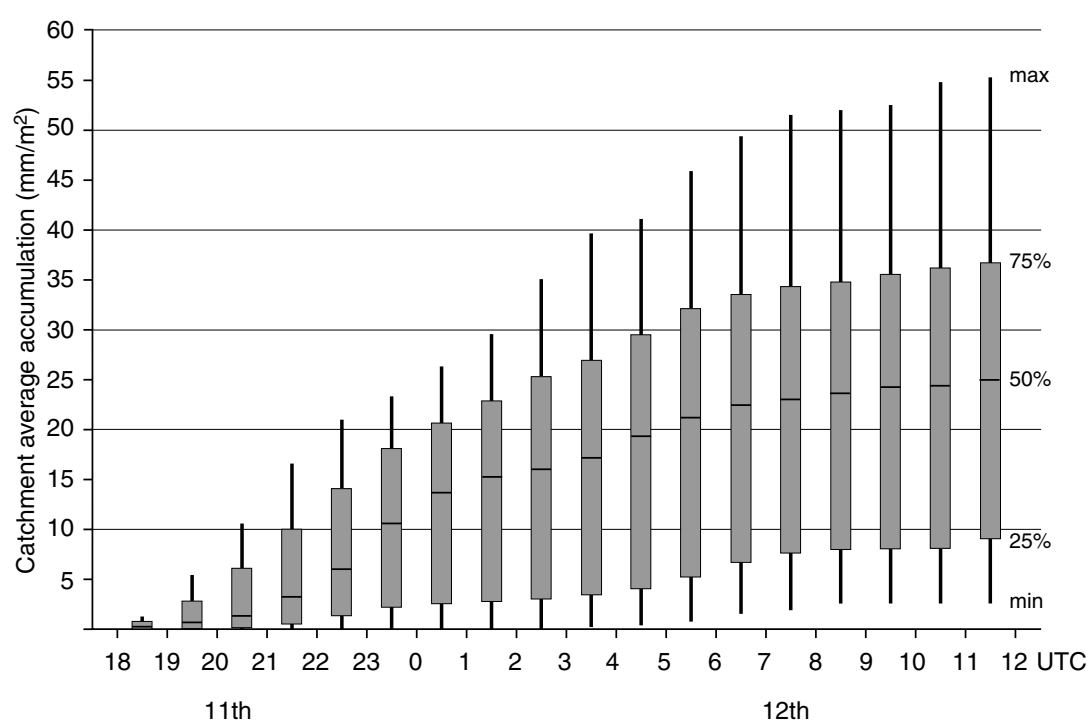


Figure 18. Graph of the cumulative catchment-average rainfall accumulations for a different catchment to that in Figure 17.

The use of an ensemble of precipitation forecast realisations as input into a rainfall-runoff model has already been examined as part of the development of the STEPS stochastic precipitation nowcast system (Pierce C et al 2004) at the Joint Centre for Hydro-Meteorological Research (JCHMR) (SPEPS is also mentioned in section 2). An example is shown in Figure 19. In this example an ensemble of 100 time-series of rainfall forecasts (at 15-minute intervals) have been fed into the PDM rainfall-runoff model (Moore, 1985, 1999; CEH Wallingford, 2001) for a particular river catchment.

The graph of multiple flow forecasts is possible because STEPS is designed to produce an ensemble of precipitation forecasts. A storm-scale model on the other hand will only produce a single forecast realisation (for the foreseeable future until computer resources allow an ensemble approach), but we have already seen that different forecast scenarios can be generated using post-processing techniques, and it is perfectly reasonable to use these in the same way.

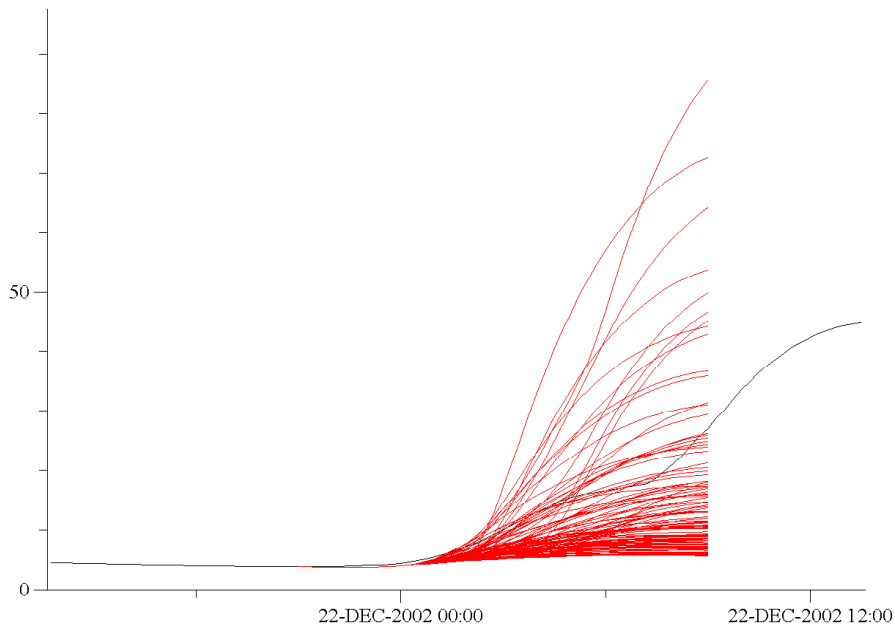


Figure 19. Ensemble flow forecasts for the River Mole made at a time origin of 21:00 21 December 2002 for a lead time of 9 hours, using 6 hours of forecast rain and a further 3 hours of zero rain. The single extended (black) line is the flow to be forecast. The Mole, draining an area of 142 km² to the gauging station at Kinnersley Manor, is a southern tributary of the River Thames on its way through London. Courtesy of Clive Pierce.

Figure 20 gives a schematic example of what this might look like for different catchment-average rainfall scenarios. The flow from the predicted time-series of catchment-average accumulations is drawn along with three other possible scenarios. The scenarios could be, for example, the 75th, 95th percentile accumulations and ‘worst-case’ accumulations given a particular forecast uncertainty. A number of different percentile curves could be obtained along with worst-case scenarios computed using several different forecast uncertainties. The system could also be automated to warn of scenarios that exceed some critical threshold for a particular catchment.

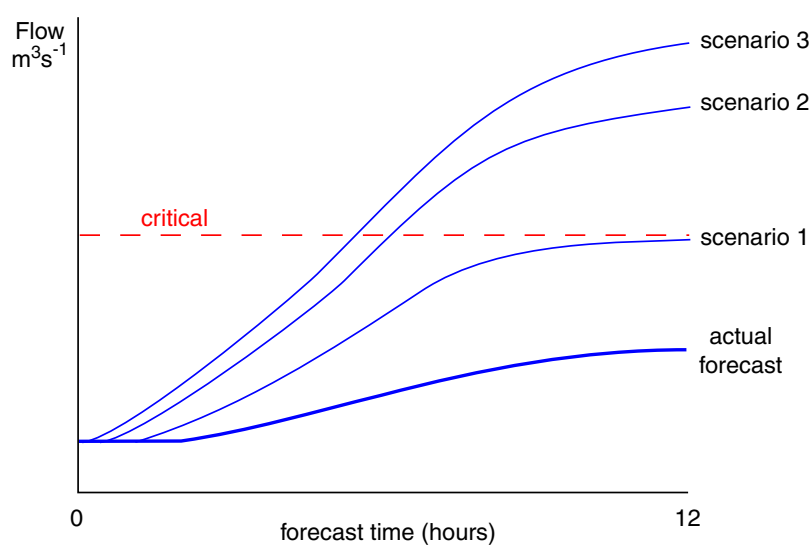


Figure 20. An example of a graph of river flow forecasts from different rainfall forecast scenarios. The red dashed line is some hypothetical critical value. See text for more information.

It need not stop there. These ensemble-style products have concentrated on larger catchment-mean accumulations, but it is conceivable that they could be generated for smaller and smaller catchments down to the size of a model grid square, provided that the uncertainty at that scale is very strongly emphasised and the forecast spread is, by definition large.

The advantage of a model with a grid spacing of 1 km (or less) is that the rainfall within a grid square can be regarded as giving a reasonable indication of what a single point measurement from a rain gauge might measure. We have confidence that the rainfall amounts produced by a 1-km model are physically sensible even though we have much less confidence in the spatial distribution of the rainfall. This means that we are getting down towards the resolution at which information can be supplied to a hydrological model as pseudo point values at specific locations rather than as catchment means (the big caveat remains the spatial uncertainty). The benefit of doing this is that it may be a more appropriate input for run-off models that have been calibrated on rain-gauge measurements. At present, hydrological models are calibrated against gauges instead of radar because certain characteristics of radar cause difficulties. The radar climatology was observed to vary according to meteorological scenarios and as post-processing changed (Moore et al). Appropriate calibration methods do need to be developed for high-resolution model output.

5.4 General comments

This section has provided examples of products that could be generated from a storm-scale modelling system for general flood forecasting and more specifically for input into hydrological models. Most of the products are designed to account for uncertainty in a forecast by generating different forecast scenarios. This is a sensible thing to do because of the stochastic nature of unpredictable small scales and the uncertainty of the larger scales. It is possible to have products that are specific to geographical areas of interest such as small river catchments because of the high density of grid points.

All of the products that have been demonstrated in section 5.2 can be computed very quickly and could be incorporated into an automated warning system. They could also be computed from more than one forecast (i.e. use not only the latest forecast but also previous forecasts).

The graphical products in section 5.3 might be very much more costly to generate if computed for a large number of river catchments. However, they need not be produced for every catchment area, only the ones for which other products had indicated that a warning level had been reached.

An area that now has considerable potential is the blending of NWP forecasts with output from nowcasting systems. Current operational practise is to blend the 12-km grid-length mesoscale model output with precipitation from Nimrod or Gandolf. The new STEPS nowcasting system will provide the extra dimension of an ensemble of rainfall predictions and give a forecast probability distribution. If this is blended with the coarser-resolution deterministic 12-km model, then the value of such a system is somewhat reduced (even with successful downscaling of the 12-km). A storm-scale model (or a 4-km grid-length model) however, has an equivalent (or better) resolution, and with appropriate post-processing is also capable of producing a rainfall probability distribution. Research is needed to combine these systems and hence make the best use of both.

6 How to verify the model objectively

Case studies have shown that a storm-resolving NWP model has the potential to produce better forecasts of significant rainfall events. However, the performance of NWP models can not be assessed properly by just looking at the output and judging ‘by eye’. Some kind of objective evaluation is also required if we wish to reveal systematic model behaviours. This section outlines an approach for verifying high-resolution NWP precipitation forecasts. More detail is given in the stage 4 report.

6.1 The verification problem

Traditional methods for evaluating model performance have generated statistics or scores on a gridpoint by gridpoint or observation point by observation point basis alone. That is not a viable approach here because we do not expect the model to be skilful at the grid scale (because of arguments presented in section 5) and it says nothing about skill over any other scales. The spatial scales we are interested in, and the spatial scales we verify on, should be compatible. We judge forecasts by eye over a variety of scales by saying that they are good at getting a large feature like a front in the right place, but are not quite so accurate for organised thunderstorms and poorer still for individual showers. But how can that be determined objectively?

Figure 21 gives an example of the nature of the problem. In this case the 1-km forecast produced a great deal more structure than the 12-km forecast and looks more like the radar picture. It gave a much better indication of the higher accumulations that we are most likely to be interested in. However, if the pictures are examined more closely, it is clear that the areas of higher accumulations in the 1-km forecast are not in exactly the same place as observed by radar. It is very likely that if a standard grid-square by grid-square verification were performed, the 1-km forecast might come out worse, even though we can see that it is a ‘better’ (or more useful) forecast. This is the challenge – to be able to objectively verify rainfall forecasts by the same criteria we use in this kind of subjective assessment (but hopefully without the bias in interpretation a human assessment might be prone to).

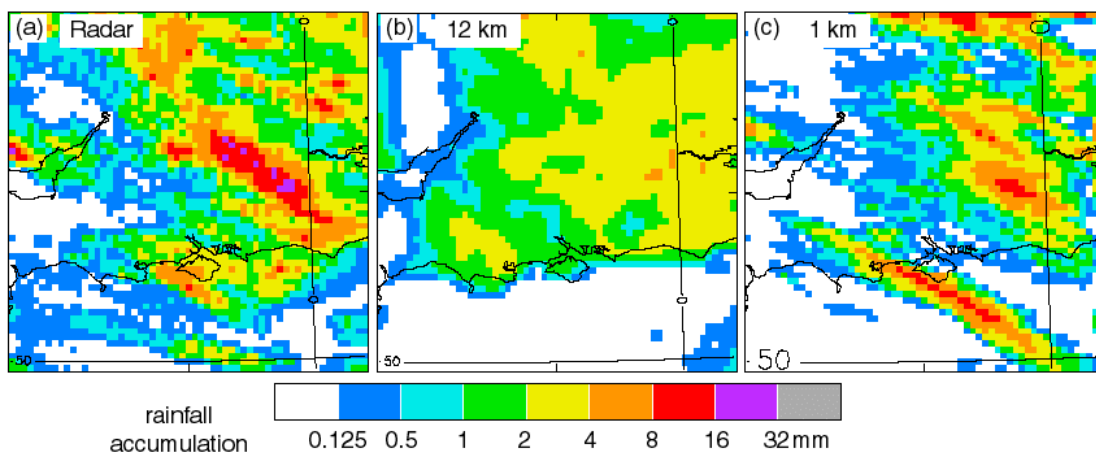


Figure 21. 6-hour rain fall accumulations for the period 10 to 16 UTC 13th May 2003 from (a) radar, (b) 12-km gridlength forecast starting at 09 UTC interpolated to the radar grid, (c) 1-km gridlength forecast starting at 09 UTC averaged to the radar grid.

6.2 Verification questions

The questions we would like to be able to at least attempt to answer in an objective way for precipitation forecasts from a storm-scale modelling system are:

1. How accurately can we forecast precipitation over areas the size of counties, river catchments or urban areas using a particular model?
2. How does the predictable scale change with forecast time? This may be in terms of defining the smallest river-catchment area for which forecasts are useful.
3. Does the rainfall analysis agree with the radar picture at the scale of the data assimilation? Data assimilation methods are designed to add observational information to a model over particular spatial scales.
4. What are sensible products to generate for customers?
5. Does a change to either model resolution or formulation make a difference to the predictable scale?

This list of questions justifies the need to have a verification system that can be used to determine the relationship between forecast skill and spatial scale. A approach for doing that has been developed and will now be outlined. It is described much more fully in the stage 4 report.

6.3 A verification method – the comparison of fractions

6.3.1 Basic decisions

Before considering the issue of spatial scale, three other more basic decisions had to be made about the way the verification was to be performed. These are listed below.

1. Verify precipitation accumulations rather than precipitation rates. The main reasons for doing this are:-
 - (a) In terms of forecasting significant events, it is rainfall accumulations that matter most.
 - (b) It is sensible to smooth in time if we also intend to smooth spatially.
 - (c) The predictability problem is reduced.
2. Verify rainfall against radar (we use analyses from the Nimrod system – Golding 1998) rather than against rain gauges. Radar data gives much better spatial coverage than rain gauges even if it is regarded as less accurate. It is available at a resolution that is comparable to that of the model, and is therefore very suitable for a like for like comparison. In short, the benefits outweigh the disadvantages (although a blend of radar and gauges may be the ideal solution).
3. Verify using accumulation exceedance thresholds (e.g. > 2 mm, > 4 mm etc) rather than accumulation amounts, so as to avoid the difficulties with a mixed distribution of zeroes and non-zero amounts.

6.3.2 Spatial scales - Computing fractions over different sized areas

For every grid square, we compute the fraction of surrounding grid-squares within a given area that exceed a particular accumulation threshold over a given period. This will give a

fraction for every grid square. The fractions can be considered as probabilities. They give an indication of the chance of an accumulation threshold being exceeded at each grid square, given that we think the model could be in error on a scale of the size of the area used to produce the fractions. It is the same approach as that used to produce the probability picture (Figure 16(a)) in section 5 (but with squares instead of circles).

Fractions/probabilities can be generated over different spatial scales by changing the size of the area. For the purposes of verification, squares of different sizes are used to compute fractions for different spatial scales. The fractions generated from identical processing of model forecast accumulations and the observed radar accumulations can then be compared.

Figure 22 gives an example of how fractions are computed over different sized squares for a particular rainfall accumulation threshold. The threshold has been exceeded where the grid squares are red and not reached where the grid squares are white. In this example, fractions are computed for the grid square marked with the black cross using two different sized squares A and B.

The fraction at grid-square X computed over square A (3x3 squares) is $6/9 = 0.667$

The fraction at grid-square X computed over square B (11x11 squares) is $66/121 = 0.545$

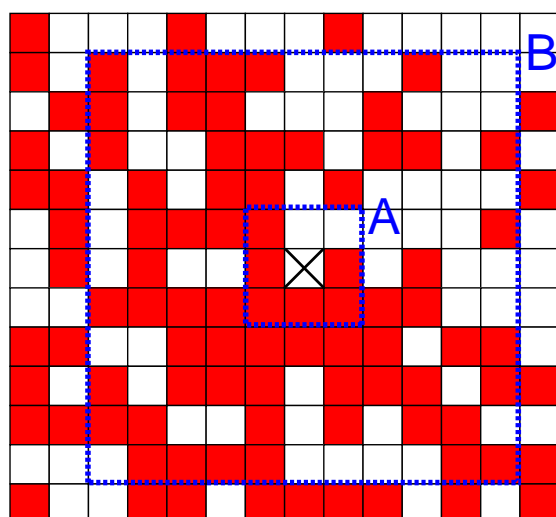


Figure 22. A schematic to show how fractions were obtained for each accumulation threshold. See text for the explanation.

The next three figures show how this looks for a particular example of a 1-km forecast and the verifying radar (the case shown in Figure 21). Figure 23 shows the pixels that exceed a threshold of 4 mm in both the forecast and radar (both on the same grid with 5x5 km pixels). Figure 24 shows the fractions/probabilities generated at every grid point (except around the edge) using the approach described for squares of size 35x35 km. Figure 25 shows the result of computing the fractions/probabilities over a larger spatial scale by using squares of size 75x75 km. At this larger spatial scale the pictures are looking more similar over a significant part of the domain. It gives the impression that there is more skill over the larger spatial scale, but also a loss of definition (or 'resolution') as the fractions/probabilities become smoother.

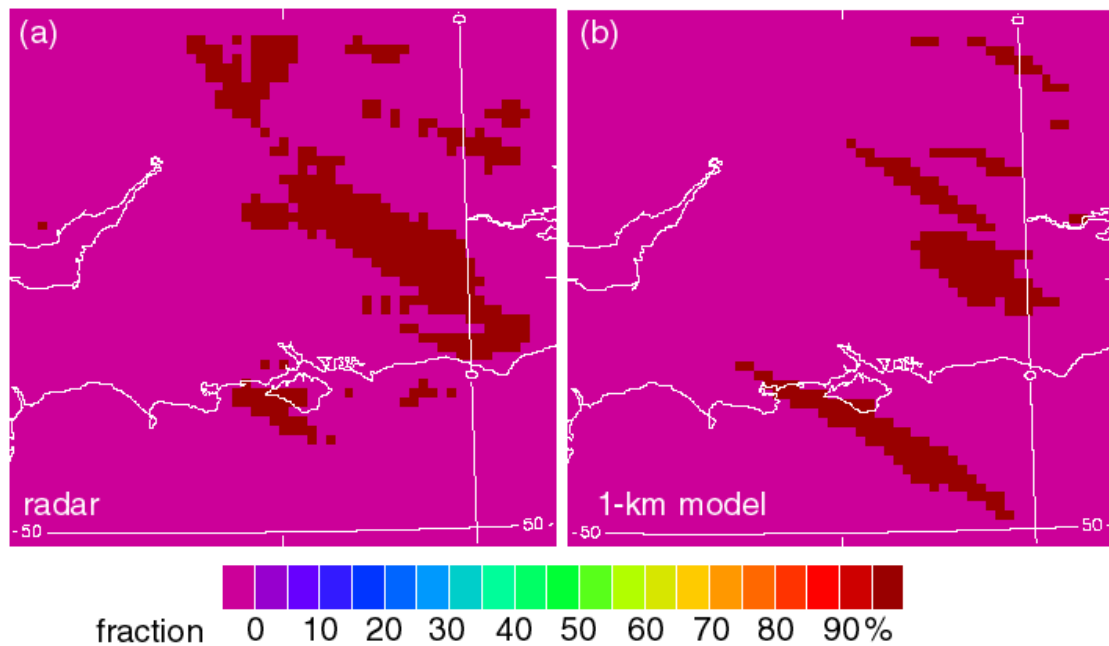


Figure 23. The pixels with rainfall accumulations exceeding 4 mm (shaded brown) from the radar and 1-km forecast 6-hour accumulations shown in Figure 21.

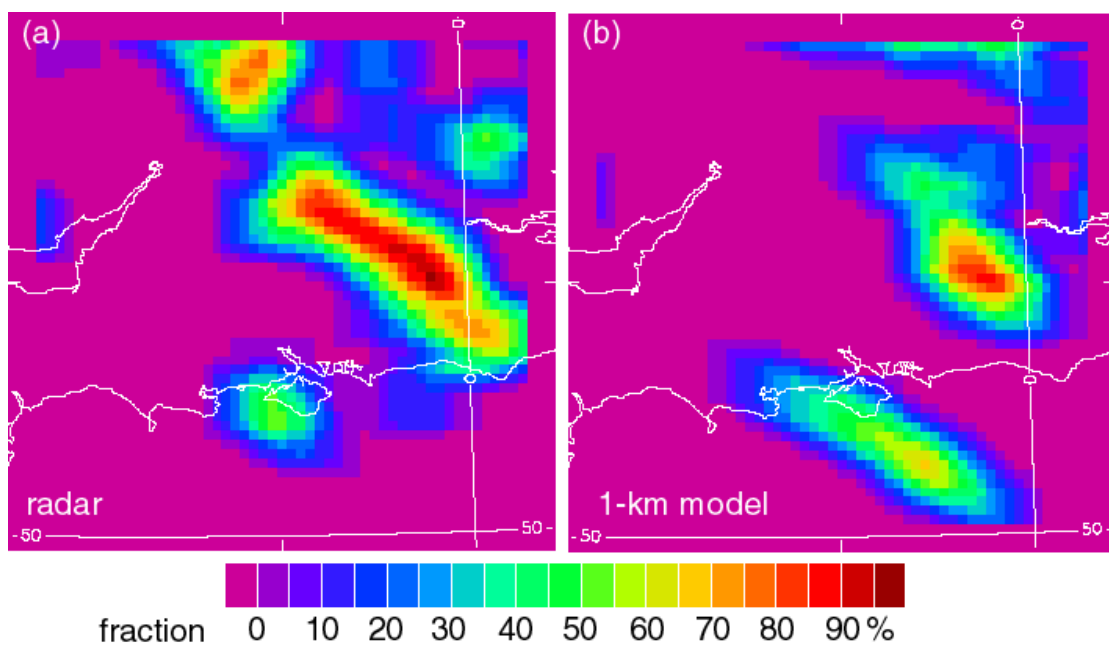


Figure 24. The fractions (or probabilities) of rainfall accumulations at each pixel exceeding 4mm, from the accumulations shown in Figure 21, using squares of 35x35 km to compute the fractions.

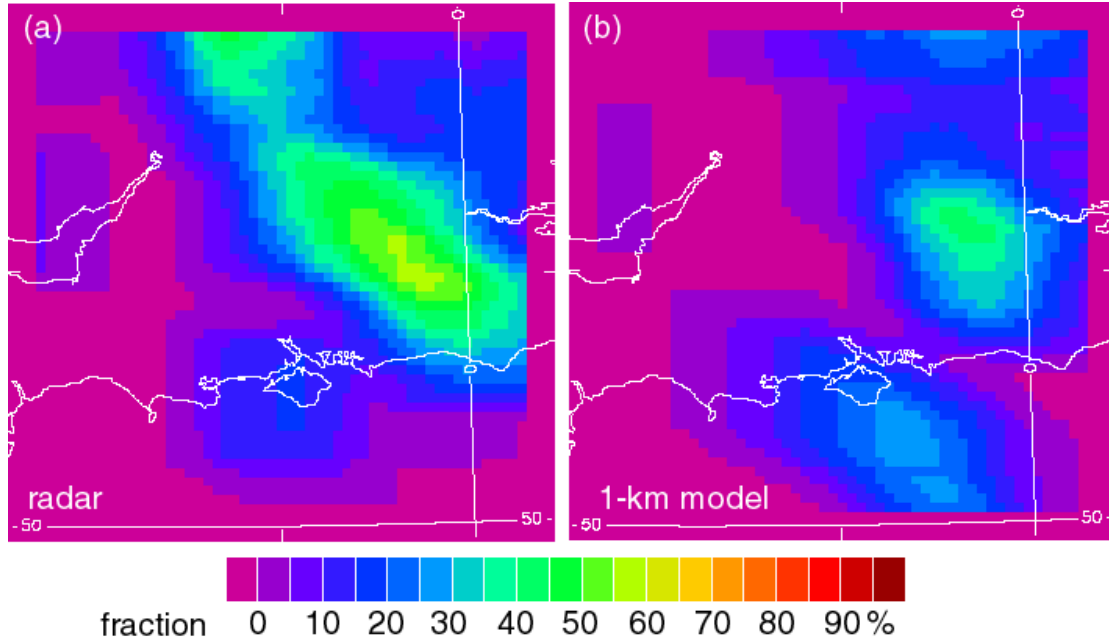


Figure 25. The fractions (or probabilities) of rainfall accumulations at each pixel exceeding 4mm, from the accumulations shown in Figure 21, using squares of 75x75 km to compute the fractions.

6.3.3 A verification score to compare fractions

There are two ways of verifying the forecast fractions for each accumulation threshold: -

1. Compare the forecast fractions with radar fractions computed in exactly the same way (as shown in Figure 24 and Figure 25).
2. Compare the forecast fractions with binary values of 0 and 1 from radar (Figure 23(a)). 1 where the threshold is exceeded, 0 where it is not.

Only option 1 is discussed here – comparing fractions with fractions.

One type of verification score used to do this has been called the Fractions Skill Score (FSS) and is a variation on the Brier Skill Score. It is given by: -

$$\text{FSS} = 1 - \frac{\text{FBS}}{\frac{1}{N} \left[\sum_{j=1}^N (p_j)^2 + \sum_{j=1}^N (o_j)^2 \right]}$$

$0 < p_j < 1$ forecast fraction
 $0 < o_j < 1$ radar fraction

where

$$\text{FBS (Fractions Brier Score)} = \frac{1}{N} \sum_{j=1}^N (p_j - o_j)^2$$

is a version of the Brier score in which fractions are compared with fractions

and

$$\frac{1}{N} \left[\sum_{j=1}^N (p_j)^2 + \sum_{j=1}^N (o_j)^2 \right]$$

is the worst possible FBS in which there is no colocation of non-zero fractions

The Fractions Skill Score has the following characteristics

- It has a range of 0 to 1, 0 for a complete forecast mismatch, 1 for a perfect forecast.
- If either there are no events forecast and some occur, or some occur and none are forecast the score is always 0.
- As the size of the squares used to compute the fractions gets larger, the score will asymptote to a value that depends on the ratio between the forecast and observed frequencies of the event. I.e. the closer the asymptotic value is to 1, the smaller the forecast bias.
- The score is most sensitive to rare events (or for small rain areas).

As with any verification score, this one has characteristics that are both helpful and misleading. It is certainly not presented here as the only way of comparing fractions with fractions, but it has proved useful for providing the sort of information we are interested in.

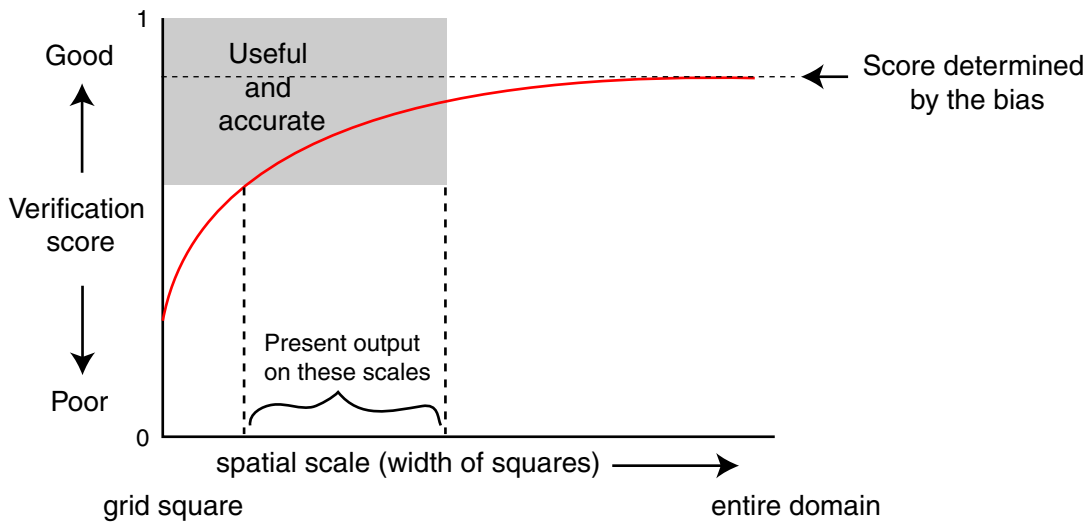


Figure 26. Schematic graph of verification score against spatial scale. See text for information.

Figure 26 shows how the Fractions Skill Score is expected to behave over a large number of forecasts for a particular accumulation period and threshold. The least skill is expected to be at the grid scale. Skill should increase with spatial scale (square size) until it reaches an asymptote that is determined by the forecast bias. The grey shading depicts the part of the graph where the score is deemed high enough for the forecast to be regarded as skilful and the spatial scale is small enough for the forecast to be useful. (There is little to be gained from a forecast that is either detailed but inaccurate or broadly accurate but lacking useful detail). We can then, in principle, pick out the range of spatial scales over which the forecast should be presented to users/customers – i.e. the size of squares used to generate probabilities.

7 The skill of the operational mesoscale model analyses and forecasts during 2003

This section stands alone somewhat as it is focussed on verification of the operational mesoscale model forecast system (grid spacing of 12 km), rather than a storm scale model. For that reason it is not critical in terms of the final conclusions of the report. However, the results are relevant for determining how well operational data assimilation methods, which are now being tested at high resolution, are able to fit rainfall forecasts and analyses to radar over a long period. It also provides a baseline for assessing high-resolution model performance in future.

The objective was to use the scale selective verification approach outlined in the previous section to investigate the ability of data assimilation methods to fit rainfall analyses to radar at different spatial scales and retain that skill into subsequent forecasts.

7.1 Outline of the investigation

The intention was to compare mesoscale model precipitation output with radar for every forecast starting from 00 and 12 UTC during 2003. In practise there were occasions when either the model forecast data or the radar data were not easily available and this meant that, in the interests of time, some forecasts were not included since the sample size remained large. The scale-selective verification was performed for hourly precipitation accumulations over the first 24 hours of each forecast. The area used is shown in Figure 27. The radar data was averaged to the same grid as the msoscale model to give a like for like comparison. Scotland was excluded because the radar data is less reliable in mountainous areas (although the same is true of parts of Wales). The only quality control of the radar performed was that done by the Nimrod processing system. The radar data was regarded as 'truth' for the purposes of this investigation.

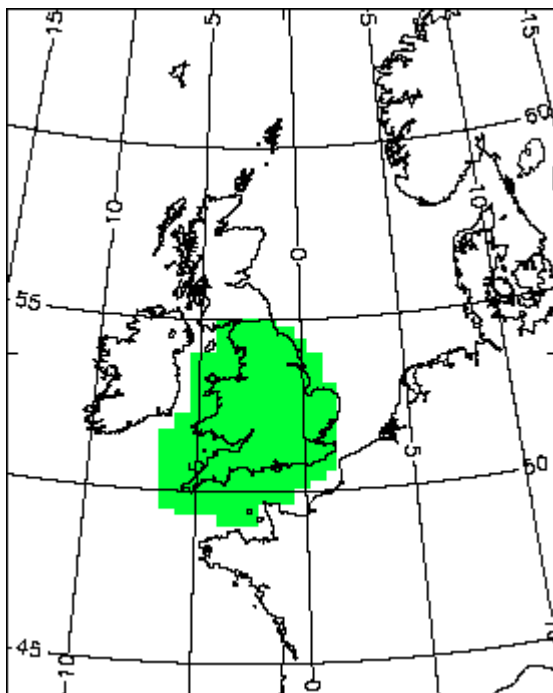


Figure 27. The green shading shows the area used for the verification of Mesoscale model precipitation analyses and forecasts.

7.2 The bias in rainfall amounts

As well as assessing the spatial accuracy of the analyses and forecasts over different scales, it was necessary to assess the accuracy of the predicted rainfall amounts. It is possible to forecast rainfall in the correct place over some spatial area but nevertheless generate too much or too little rain. In other words, to get a full picture, we also need to know about any bias in the precipitation forecasts over the whole verification area. For that reason the average amount of rain produced by the model on an hour by hour basis has been compared with the equivalent quantities measured by radar. Figure 28 shows this comparison for all the forecasts examined and for a spring/summer period when convective precipitation made a significant contribution.

Results from the whole year

See graphs (a), (b) and (c) (the first column) in Figure 28. There seems to be a double peak in the graph of the average amount of rainfall measured by radar over the year against forecast time when both the 00 and 12 UTC forecasts are assessed together (indicated by the two 'P's in (a)). A separation of results from the 00 and 12 UTC forecasts (b & c) reveals that there was a peak in the amount of rainfall around 06 UTC. At first sight this seems strange; it would be more reasonable to expect very little variation in rainfall amount over a day when averaged over a whole year. If there was to be peak, intuitive reasoning might suggest that it should be during the late afternoon when convection is most active. A closer examination of the month by month values (not shown) showed that much of this signal comes from the very wet January when most of the rain did indeed fall around 06 UTC. In fact, 2003 was a peculiar year, in that there were some extremely wet periods with repeated daily patterns and an unusually dry hot summer. The mesoscale model did successfully produce a peak in the morning (b) and (c), but with a delay of a few hours. In general the model produced mean rainfall amounts that were in good agreement with the radar (within 15% of the radar value at worst). However, at the start of the forecasts, the model produced considerably too much rain (Labelled A in (a), (b) and (c)). This was true in both the 00 and 12 UTC forecasts. They produced 40% too much rain on average over the first 2 hours, indicating that there is a problem with the data assimilation. The most likely cause is the MOPS latent heat nudging, which modifies the rainfall even in the second hour of the forecast.

Results from April to August

In order to isolate or remove the signal from the very wet January, the mean hourly accumulations from different parts of the year were also examined. The graphs from the period April to August are presented in (d) to (f) (second column) to show results from the period when convection was most active over land. They also show a double peak even without the January data. The difference is that over that period, there was a peak around 18 UTC as well as around 06 UTC. The peak at 18 UTC can be explained by an increase in convective activity during the late afternoon/evening. The continued peak around 06 UTC was more of a mystery, but inspection of the hour by hour radar pictures has shown that much of the rain fell in a wet period from the final third of April through to the middle of May and that (like in January) a disproportionate amount of that rain fell during the morning around 06 UTC. So, in fact, due to the quiriness of rainfall occurrence in 2003 there was also a peak in the morning even in the Spring/Summer sample, but that is reassuring as it is thought to be largely real and not an artefact of the radar.

The mesoscale model has the same problem of over prediction at the start of the forecast during April to August as seen over the year as a whole. In addition another deficiency can be detected. The letter 'M' in (e) and (f) has been added to indicate that the model

produced too much rain on average in the late morning. The letter 'E' (also in (e) and (f)) has been added to show that the model produced too little rain on average in the evening. Both the under and over prediction are linked to the behaviour of the convection parametrization scheme. The convection scheme triggered too early in the day, which led to the over prediction in the morning, then because it was unable to organise convection, stopped triggering too early as the solar heating decreased in the afternoon, and this led to the under prediction in the evening. A storm scale model should not have this problem because convection parametrization is not needed.

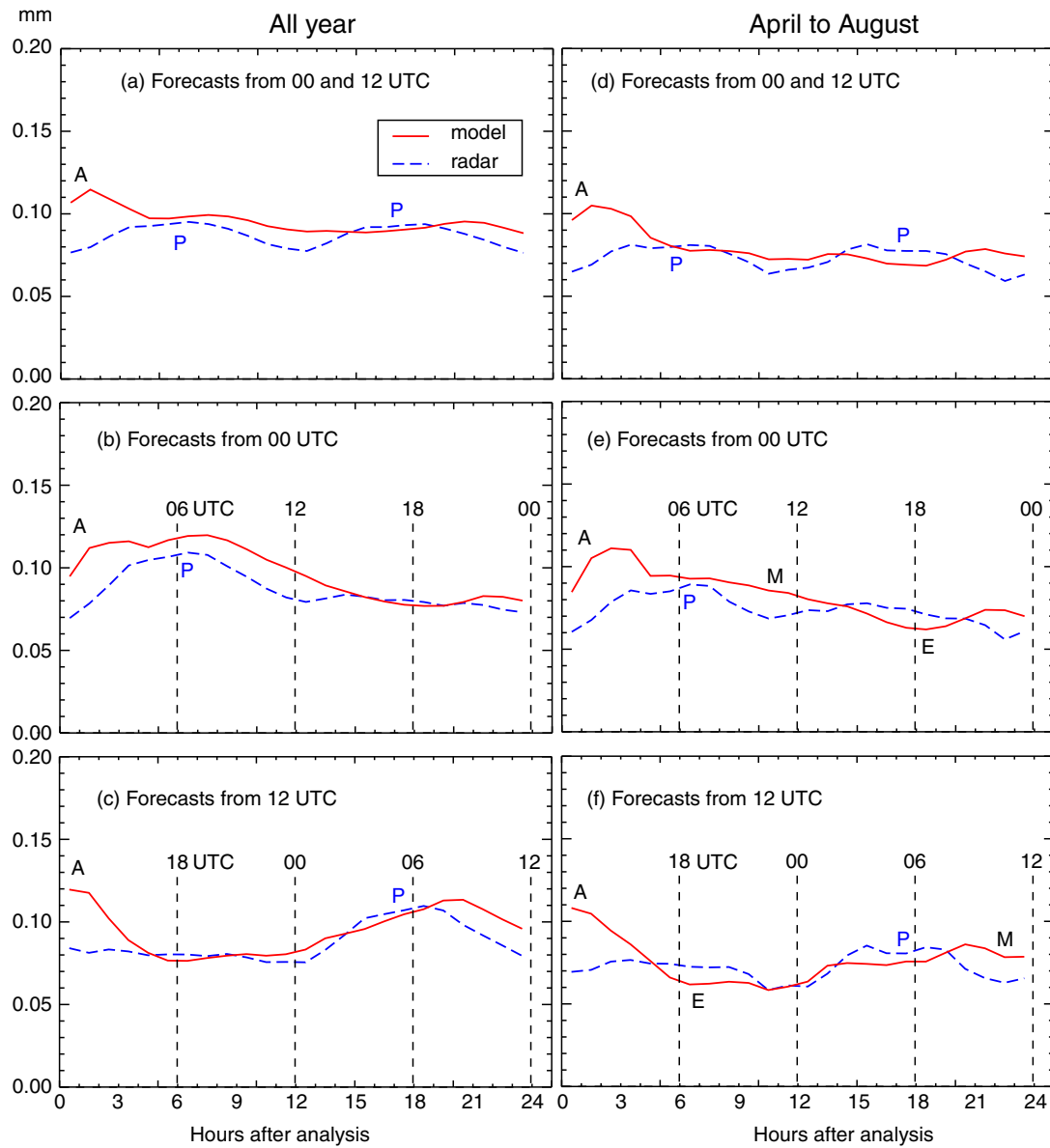


Figure 28. Graphs of mean hourly rainfall accumulations from radar (blue dashed lines) and the mesoscale model (red lines) over the verification area (green in Figure 27). Graphs on the left - Results from the whole of 2003 (a) forecasts from 00 and 12 UTC, (b) forecasts from 00 UTC, (c) forecasts from 12 UTC. Graphs on the right - Results from April to August 2003 (a) forecasts from 00 and 12 UTC, (b) forecasts from 00 UTC, (c) forecasts from 12 UTC. Letters 'A', 'P', 'E' and 'M' are discussed in the text.

7.3 Scale-selective skill of the spatial distribution of the precipitation forecasts

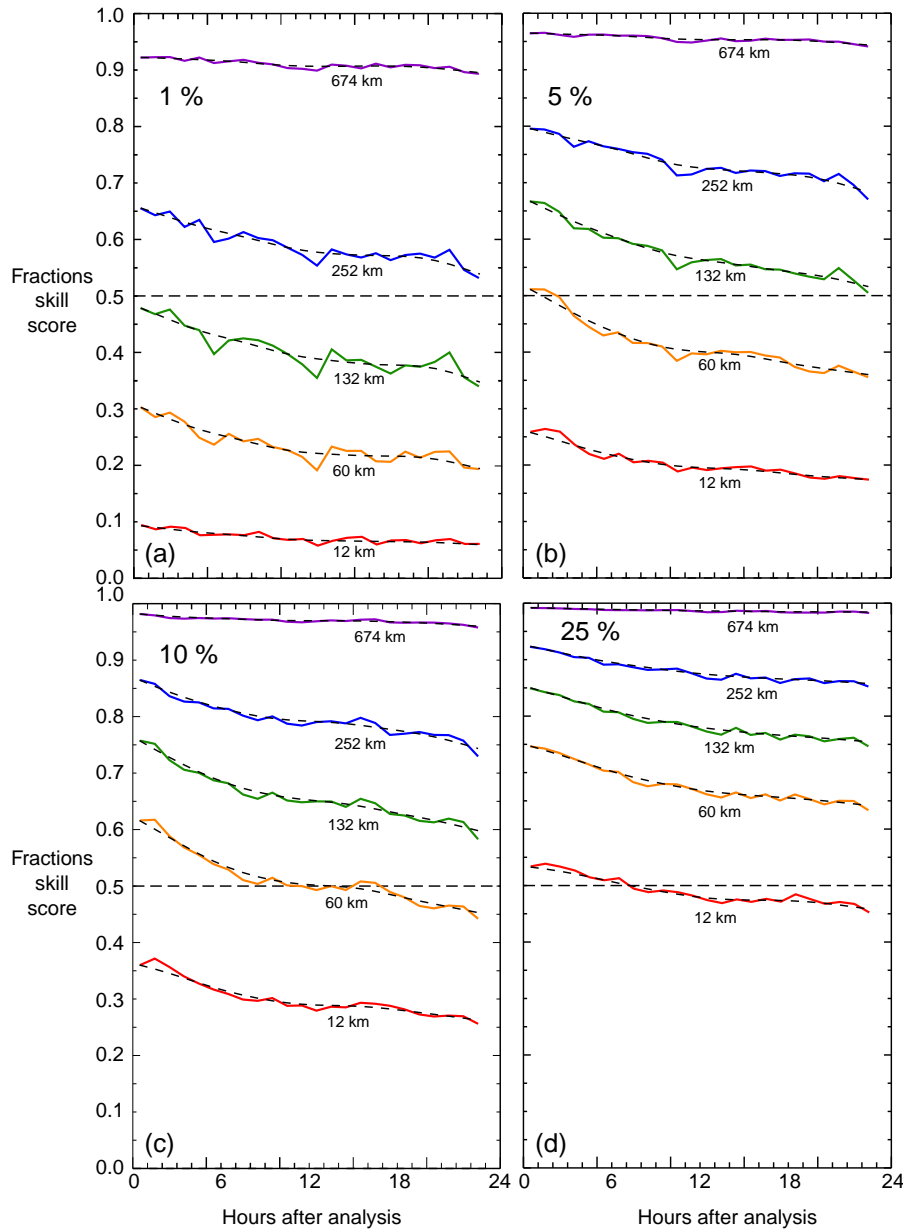


Figure 29. Graphs of Fractions Skill Scores against forecast time for forecast hourly accumulations compared with radar over different spatial scales (fractions from squares of length 12 km (grid scale), 60 km, 132 km, 252 km and 674 km). Thresholds are the top (a) 1%, (b) 5%, (c) 10% and (d) 25% of hourly accumulations from all grid squares within the verification area. The dashed lines show smoothed versions of the lines using 5 iterations of a 3-point mean.

Figure 29 shows results from the calculation of the Fractions Skill Score (described in the previous section) for the comparison of fractions from the mesoscale model against radar for five different sized squares (spatial scales) and using four different thresholds.

The four thresholds used were the top 1%, 5%, 10% and 25% of hourly accumulations (i.e. the 99th, 95th, 90th and 75th percentile hourly accumulation values). I.e. for the 1%

threshold the locations of the top 1% of forecast accumulations was compared with the top 1% of radar accumulations. This meant that the actual accumulation thresholds might be quite different if the model had a bias, e.g. say 2mm for the forecast data and say 5mm for the radar data. The accumulation thresholds were recomputed for each comparison and the combined scores only generated from occasions when there had been sufficient rain for the threshold accumulation to be greater than zero in both model and radar. The reason for choosing this type of threshold instead of actual accumulations (e.g. 2 mm) was to remove the bias (because by definition the number of model and radar pixels is made the same) in order to concentrate on the spatial distribution of the rainfall forecasts. The bias was dealt with earlier.

The key results are:

1. For all the thresholds the skill in forecasting the distribution of precipitation increased with spatial scale. In other words, the forecast fractions (or probabilities) were closer to the radar fractions over larger areas.
2. The model was more skilful at predicting more widespread areas of rain (25% threshold) and least skilful for the more isolated events or higher accumulations (1% threshold).
3. The forecasts were most accurate at the start, indicating that data assimilation was capable of improving the accuracy of the rainfall distribution. The amount of improvement depended on the threshold and spatial scale. This variation is discussed later.
4. The skill of the forecasts dropped with forecast time. The way it dropped depended on the threshold and spatial scale. This variation is discussed later.
5. The information in this graph could be used to determine the smallest spatial scales over which to present model output. For example, if we take the 5% threshold and consider a score of 0.5 to be acceptable then output products should represent scales of around 60 km at the start of the forecast and around 130 km after 24 hours.

7.4 The scales over which the data assimilation operated

The results from Figure 29 have shown that data assimilation has produced a better fit of the predicted rainfall distribution to radar at the start of forecasts (analysis time) (point 3 above). The size of the impact of the data assimilation can be seen by measuring the difference in skill between the analysis time and a later forecast time. If this is done using the Fractions Skill Score as the measure then the impact of the data assimilation over different spatial scales can also be examined.

Figure 30 shows the difference between Fractions Skill Scores from the smoothed dashed lines in Figure 29 between the first hour and the sixth hour of the forecasts (comparison of the first hour with other forecast times gave very similar results). It provides an indication of how much extra skill in rainfall distribution the data assimilation has added to the start of the forecast, and over what scales. The curves in Figure 30 reveal that most of the extra skill was added over spatial scales of between 40 and 150 km (the peaks in the curves). This result was expected as it agrees with the scales over which we would expect the data assimilation to operate most strongly. The correlation length scale for 3DVAR in the mesoscale model is ~90 km. MOPS operates at scales > 20 km. Much less skill was added at the grid scale or over very large scales because the data assimilation methods are not designed to influence those scales so much. The differences between the curves from the different thresholds are interesting, but any explanation would be speculative and is not presented here.

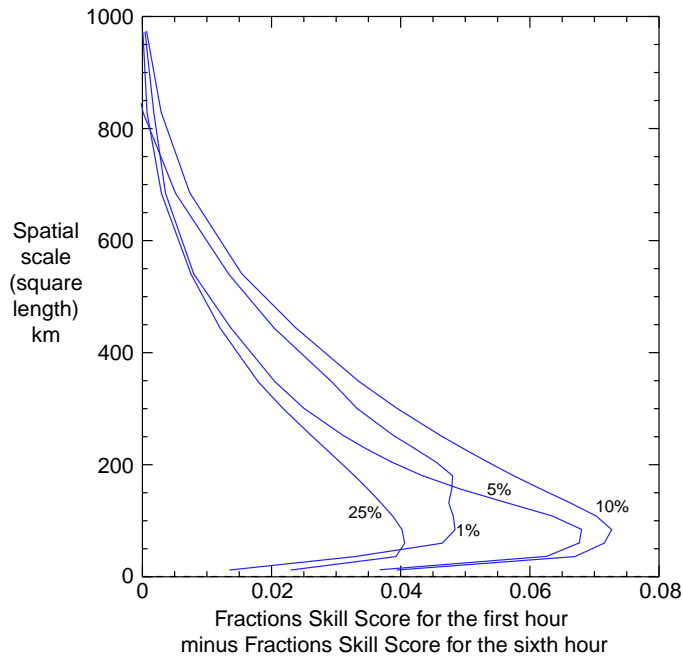


Figure 30. Graph of the Fractions Skill Score for the first hour minus the Fractions Skill Score for the sixth hour against spatial scale for accumulation thresholds representing the top 1%, 5%, 10% and 25% of grid squares over the verification area.

7.5 The retention of skill

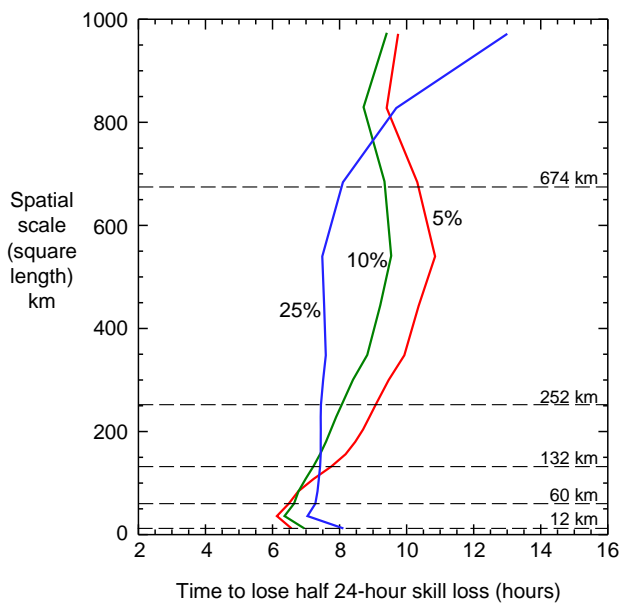


Figure 31. Graph of the timescale for skill loss against spatial scale for the 5, 10 and 25% accumulation thresholds. See text for more information.

We have seen from Figure 30 that most skill was added by the data assimilation system at scales between 40 and 150 km. However, it does not give any information about how well the skill was retained into the forecasts over different spatial scales. Figure 31 shows how

well the skill was retained. The graph shows the time it took for the model to lose 50% of the difference in skill between the first hour and the final (24th) hour. The first thing to notice is that for the 5% and 10% thresholds, skill was retained longer as the spatial scale increased and retained longest at around 550 km. At scales longer than 550 km, skill retention again decreased, although the values are less reliable at the longest scales. For the 25% threshold, the retention of skill was almost the same at scales between 60 and 550 km. All the thresholds show the most rapid decrease in skill at 36 km (3 model gridlengths) rather than at the grid scale. The explanation for this behaviour may be that 36 km is the point on the graph closest to the scale at which MOPS latent heat nudging is designed to operate. The latent heat nudging managed to improve the analysis more at that scale than at the grid scale (as seen in Figure 30), but much of this improvement was not as well retained.

7.6 Main conclusions

- The start of the forecasts showed the best fit of the distribution of precipitation to that observed by radar. Data assimilation was able to improve the rainfall distribution.
- There was 40% too much rain in the model at the start of the forecasts. This error became less with forecast time after the second hour. Data assimilation introduced a bias into the amount of precipitation. The maintenance of the bias into the second hour suggests that MOPS latent heat nudging may be the cause.
- The data assimilation added most skill to the distribution of rainfall at scales of between 40 and 150 km but retained the improvement in skill for longer over scales larger than 150 km, peaking at or greater than around 550 km.
- The loss of skill following the analysis was most rapid at a spatial scale of around 30-40 km rather than at the grid scale. This may be a deficiency of MOPS latent heat nudging. In contrast, the skill at the grid scale was retained longer than at 30-40 km, but much less skill was added.
- The data assimilation only resulted in improved forecasts at scales well resolved by both data and the model.
- These results are considered reliable enough to be used as a baseline for comparison with a large sample of forecasts from storm-scale model. A note of caution is that 2003 was a quirky year for rainfall and perhaps it would be worthwhile to add another year to the results.

8 Storm-scale model performance scores

This section provides the key results from an objective assessment of high-resolution forecasts. Objective performance scores were computed using the approach outlined in section 6. They were obtained from 7-hour precipitation forecasts over four separate days in spring/ summer 2003 using grid-spacings of 12, 4 and 1 km. This section gives a summary of the key results, more detail is provided in the stage 5 interim report. The four days were chosen purely because significant convection occurred in the HRTM 1-km model domain (see Figure 2) and they were used as part of an HRTM project testing phase. They were not selected on the basis of the operational 12-km model's performance. For each of the four days, forecasts were run that started at 06, 09, 12 and 15 UTC.

The days chosen were:

1. 13th May 2003
2. 25th May 2003
3. 1st July 2003
4. 28th August 2003







The aim was to examine both the impact of changing model resolution and the impact of using data assimilation at high resolution on forecast skill over different spatial scales. Verification scores were obtained for hourly and 6-hourly rainfall accumulations. The hourly accumulation scores show the variation of skill with forecast length.

Five different model configurations were examined.

- (1) 12-km gridlength with data assimilation (3DVAR and MOPS)
- (2) 4-km gridlength with additional data assimilation on the 4-km grid (3DVAR and MOPS).
- (3) 4-km gridlength with no additional data assimilation - i.e. starting exactly the same as the 12-km forecast. So called 4-km 'spin up' run.
- (4) 1-km gridlength with additional information added from part of the data assimilation at 4 km (3DVAR). So called 1-km 'added increments' run.
- (5) 1-km gridlength with no additional data assimilation - i.e. starting exactly the same as the 12-km forecast. So called 1-km 'spin up' run.

Further information about the data assimilation methods 3DVAR and MOPS is given in the stage-4 report. These configurations represented the data assimilation capability at the time the verification was performed.

The following colour/line convention is used for displaying the verification graphs presented in the remainder of this section.

	1 km gridlength	spin up from 12 km
	4 km gridlength	spin up from 12 km
	1 km gridlength	added increments from 4 km VAR assimilation
	4 km gridlength	4 km VAR assimilation + 4 km MOPS
	12 km gridlength	12 km VAR assimilation + 12 km MOPS
	radar	

8.1 Scores for 6-hourly accumulations

8.1.1 Fractions skill score

Figure 32 shows the Fractions skill scores for a selection of 6-hour rainfall accumulation thresholds. The threshold of 8 mm is the largest for which scores are presented because of the rarity of larger accumulations over the small number of events.

The key results are:

General results

1. The forecasts have most skill at predicting small rainfall accumulations. This is partly a consequence of the verification measure used.
2. As expected, the skill of all the forecasts increased with spatial scale, particularly for squares of width < 100 km. All the forecasts have least skill at the grid scale.

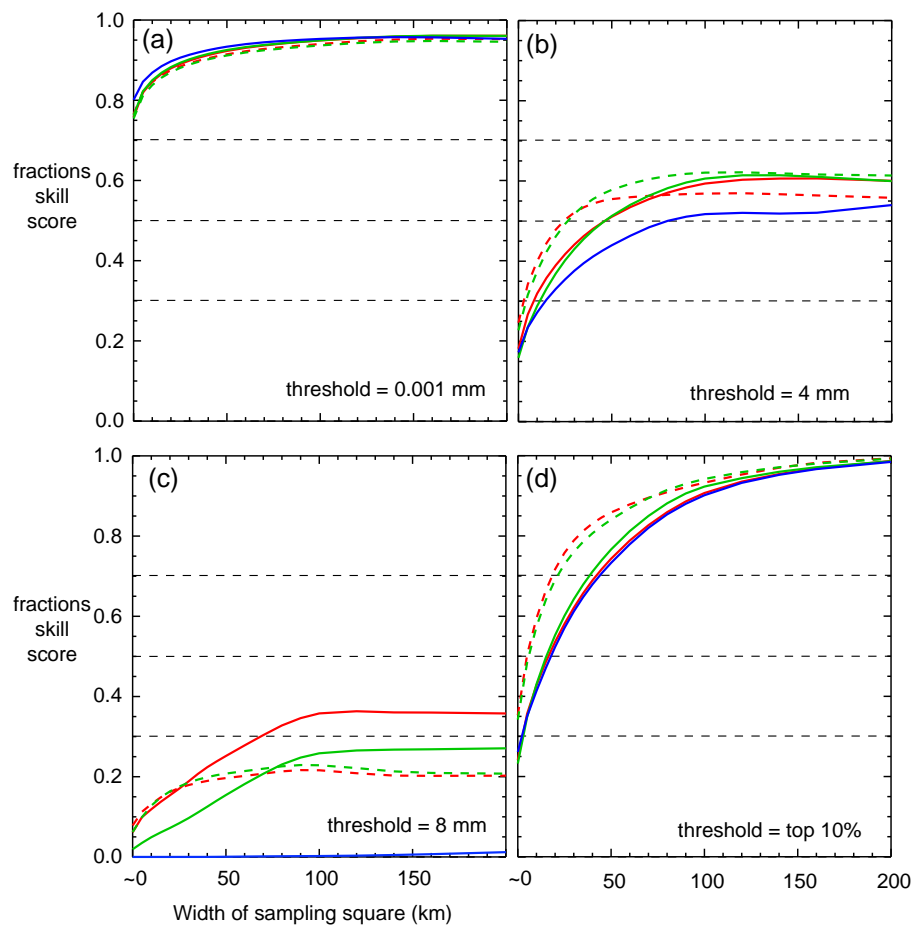


Figure 32. Fractions skill scores against square size (spatial scale) for 6-hour accumulations exceeding thresholds of (a) 0.001 mm, (b) 4 mm, (c) 8 mm, (d) 90th percentile value. The bottom axis shows the width of the sides of the squares over which the forecast and radar fractions were computed. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

Specific results:

1. The forecasts were mostly comparable for the threshold of 0.001 mm (essentially rain or no rain). The 12-km forecasts were slightly better.
2. The skill scores for the accumulation threshold of 4 mm show that over smaller spatial scales (square width < 60 km), the 'spin up' 1 and 4-km forecasts (dashed lines) were the most skilful, the 1 km with added increments (red line) and 4 km with data assimilation were similar and the next most skilful, and the 12-km forecasts were the least skilful. At larger scales (square width > 60 km), the 1-km with added increments along with both 4-km configurations were the best, and the 12 km remained the worst. At the grid scale all the forecasts had the least skill and were similar.
3. The skill scores for the accumulation threshold of 8 mm show that over small spatial scales (square width < 30 km) the 1-km forecasts with added increments and the 1 and 4-km 'spin up' forecasts were the most skilful. At larger spatial scales the 1-km forecasts with added increments were the best. The 12-km forecasts had practically no skill as they failed to produce accumulations > 8 mm on nearly all occasions.
4. When the threshold of the top 10% of grid squares (90th percentile value) is used instead of a specific accumulation threshold, the 1 and 4-km 'spin up' forecasts gave the best scores over all spatial scales. This indicates that the spatial distribution of the higher accumulations was better represented in those forecasts. All the curves approach 1 with increasing spatial scale because there is inherently no bias over the domain when using a percentile threshold.

8.1.2 Brier skill scores

The Brier skill scores are shown (Figure 33) for a threshold of the top 10% accumulations (90th percentile). They have been computed by comparing forecast fractions with binary values of 0 or 1 from radar, rather than comparing fractions with fractions as described in section 6. The score can range from 0 for a forecast with no skill to 1 for a perfect forecast, and if a percentile threshold is used, will asymptote to 0.5 for large spatial scales (equivalent to the domain size). For more information refer to the stage 4 and 5 reports. Two values are important, (1) the smallest scale at which a curve exceeds 0.5 (if at all) and (2) the spatial scale of the peak in the curve (if there is one). The purpose of using a percentile threshold rather than an accumulation threshold is to see how forecasts compare when the bias is removed.

The curves in Figure 33 show that the 1 and 4 km 'spin up' forecasts (dashed lines) were the most skilful overall. The 1-km with add increments (red line) and the 4-km with data assimilation (green line) were the next best. The 12-km forecasts were the worst. This agrees with the results obtained from the Fractions skill score using the same threshold (Figure 32(d)).

From the peaks in the curves, the most useful spatial scale for the prediction of the location of the highest 6-hour accumulations was around 30 km for the 1 and 4 km 'spin up' forecasts, 55 km for the 1 and 4 km forecasts with added data and 70 km for the 12-km forecasts.

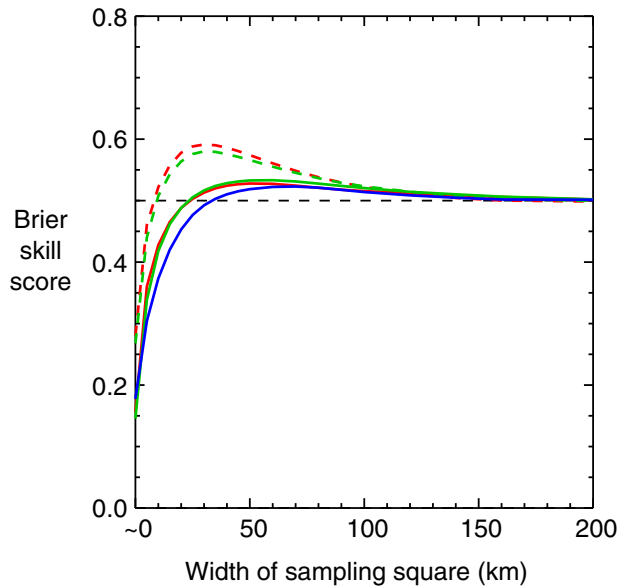


Figure 33. Brier skill scores against square size (spatial scale) for 6-hour accumulations exceeding the 90th percentile value (top 10% of accumulations within the domain). The bottom axis shows the width of the sides of the squares over which the forecast and radar fractions were computed. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

8.2 Scores for hourly accumulations

The graphs in Figure 34 show how the skill of each of the forecasts, over a particular spatial scale (square of width 25 km), changes from hour to hour within the forecast period.

The key results from the graphs in Figure 34 are:

Accumulation threshold of 0.01 mm.

- (1) The 4 and 1-km 'spin-up' forecasts were significantly worse than the others over the first 2-3 hours. The 1-km became comparable to the others after 2 hours and the 4-km after 3 hours. This is due to the time it takes for these models to develop higher-resolution structure from a lower-resolution starting point.
- (2) The 12-km forecasts were the most skilful throughout the forecast period, but not by a great margin.
- (3) There is no loss of skill through the forecast period from any of the forecasts.

Accumulation threshold of 1 mm

1. The 4 and 1-km 'spin-up' forecasts (dashed lines) were the least skilful over the first 3 hours. They became comparable to the other 4 and 1-km forecasts (green and red lines) for the rest of the forecast period. Again, this is due to the time needed to develop high-resolution structure.
2. The 12-km forecasts (blue line) and the 4-km forecasts with additional data assimilation (green line) are the most accurate at the start. The 1-km forecasts with added increments (red line) is not as good (though better than the 'spin-up' forecasts). This result may be showing a beneficial impact of MOPS on the 12 and 4-km forecasts.

3. The 12-km forecasts were the least skilful over the last 2 hours. An examination of the model output suggests that this may be a result of an inability to maintain showers into the evening in some of the forecasts.
4. There seems to be a peak in skill at 4-5 hours in all the forecasts. This may be a real signal, but is more likely to be just an artefact from sampling a small number of forecasts.

Frequency threshold of 5% (top 5%) (graph c)

This graph looks very similar to the 1 mm threshold:-

1. The 1 and 4-km 'spin-up' forecasts (dashed lines) had the poorest scores early in the period, but with this threshold became the best after 3-4 hours, until the end when they were comparable to the 1 and 4-km data assimilation forecasts.
2. The 1-km forecasts had the least skill over the last 2 hours
3. The 1 and 4-km forecasts with data assimilation (red and green lines) were very similar, with the 4-km having slightly higher scores up to 5-hours into the forecast. Again, this may be the impact of MOPS.
4. The unexplained peak in skill at 4-5 hours is just as evident as it is with a threshold of 1 mm.

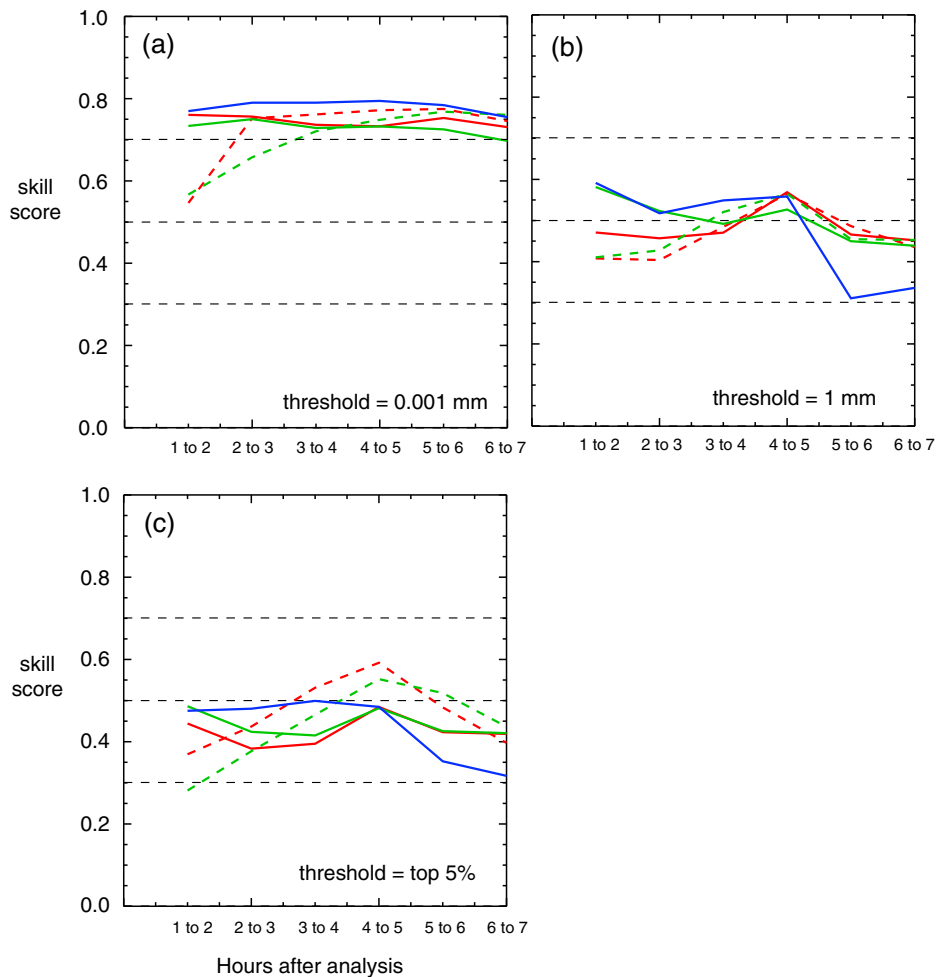


Figure 34. Fractions skill scores for hourly accumulations exceeding thresholds of (a) 0.001 mm, (b) 1 mm, and (c) 95th percentile value, within squares of size 55x55km for all four days. The bottom axis shows the 1-hour accumulation periods after the start of the forecast. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km.

The key results from the graph in Figure 35 are:

1. The 1 and 4 km ‘spin up’ forecasts (dashed lines) started off producing too little rain and ended up producing far too much. The 1-km forecasts (red dashed) generated too much rain more quickly, but the 4-km forecasts (green dashed) ‘overshot’ by more.
2. The 1 km forecasts with added increments (red line) and the 4 km forecasts with data assimilation (green line) produced too much rain at the start, became comparable to the radar in the middle of the period and had slightly too little by the end (assuming zero error in the radar).
3. The 12-km forecasts (blue line) produced the closest mean accumulation to the radar, though still too much at the start (just as evident from assessment of 12-km forecasts over 2003 – previous section).

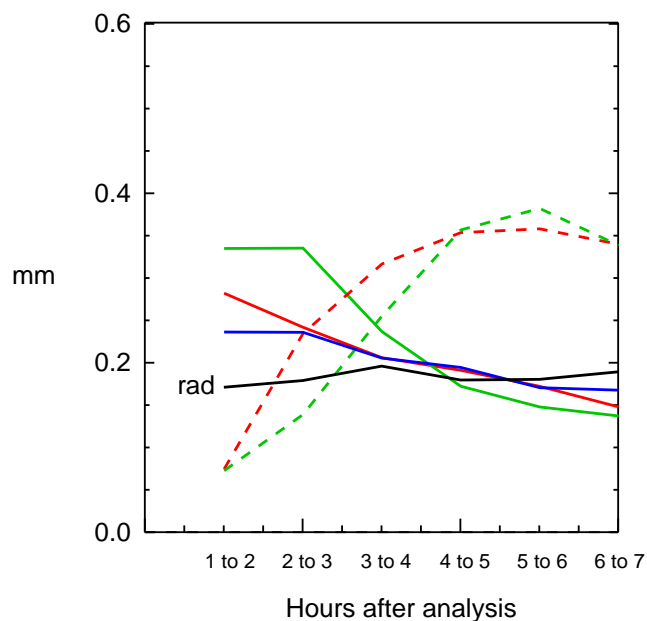


Figure 35. The mean hourly accumulations over the 1-km model domain for all four days. The bottom axis is the 1-hour accumulation periods after the start of the forecasts. Blue line = 12km forecast, red continuous line = 1km with added increments, blue continuous line = 4 km with data assimilation, red dashed line = 1 km spinning up from 12 km, blue dashed line = 4 km spinning up from 12 km, black line = radar.

8.3 Implications of the results

The verification was performed to obtain an objective and scale-selective measure of the impact of model resolution and data assimilation on precipitation-forecast skill. It was successful in meeting those objectives. Not only that, the verification scores were in agreement with a subjective assessment of the events (see the stage 5 report), which means that we can have confidence in the results. The impact of resolution and data assimilation on forecast skill will now be discussed in turn.

8.3.1 The impact of resolution

The results support the findings of the initial subjective assessment; that a model with a grid spacing of 4 km or shorter is capable of more skilful predictions of higher-accumulation rainfall events than a model with a grid spacing of 12 km. The main reason appears to be that the higher-resolution forecasts were able to generate the locally higher rainfall accumulations, though not necessarily in quite the right place. The 12-km forecasts were not able to do this. For accumulation thresholds of 4 and 8 mm, the higher resolution forecasts (1 and 4 km grid spacing) gave the best scores.

So which resolution is best? The initial subjective assessment concluded that the 1-km (or 2-km) grid-length model performed better than either the 12 or the 4km for the thunderstorm events that were examined. Can the same conclusions be drawn from these results?

If we examine the ability of the forecasts to predict the spatial distribution of the higher accumulations, the results show that the 1 and 4 km ‘spin up’ forecasts were generally the most accurate and the 1 km forecasts were slightly the better of the two. Useful accuracy was achieved by these forecasts over spatial scales of around 20-30 km, compared to around 70 km for 12 km. The problem with the ‘spin-up’ forecasts was a tendency to over predict the amount of rain once the convective cells had started to form.

It is more difficult to draw conclusions about the ability of the different resolutions to predict high accumulations associated with severe thunderstorms because none of the four days had storms of the same intensity (or longevity) as those seen in the initial case studies. This is a consequence of the lack of severe convection over southern England in the hot dry summer of 2003. However, the graph of fractions skill scores from the highest accumulation threshold of 8 mm, does show that the 1-km grid-length model with added increments was the most skilful. This was over larger spatial scales (square width > 30 km); the skill at smaller scales was poor in all forecasts.

So it does seem that a grid spacing of 1 km gave the best forecasts of the most significant convective rainfall events. The 1-km ‘spin-up’ model gave the best forecasts of rainfall distribution (though only marginally better than the 4-km spin-up), and the 1-km forecasts with ‘added increments’ gave the best predictions of the highest accumulations.

8.3.2 The impact of data assimilation

Data assimilation is necessary if we are interested in the first few hours of a high-resolution forecast. The results from the hourly accumulations back this up. They show that the 1 and 4 km ‘spin up’ forecasts (with no high-resolution data assimilation) have low skill at the start (first 2-3 hours), particularly for high accumulation thresholds.

The data assimilation methods applied here (added increments at 1 km and 3DVAR and MOPS at 4 km) have had a beneficial effect in improving the skill at the start of the forecasts, but unfortunately this was not sufficiently maintained throughout the forecasts. The high-resolution forecasts with data assimilation were not able to predict the spatial distribution of the higher accumulations as well as the high-resolution ‘spin up’ runs (though still better than 12 km). It appears that the data assimilation, by adding extra

high-resolution information, was somehow disturbing the coherence in the development of convection contained within the ‘spin up’ runs.

A positive impact of data assimilation was a reduction in the serious over-prediction of rainfall amounts that were observed after a few hours in the ‘spin-up’ forecasts. However, this was at the expense of an over-prediction at the start of the forecasts (albeit less serious). The other positive result was the improvement to the 1-km scores for the 8 mm accumulation threshold, and since it is the higher accumulations we are primarily interested in, this is encouraging.

In short, the results are mixed. They confirm that data-assimilation in a high-resolution modelling system is both necessary and difficult. There has been some success in improving the spatial distribution and intensity of rainfall at the start of the forecasts, but this has mostly not led to greater spatial accuracy at later times. However, we should remember data assimilation in high-resolution models (grid spacing < 5 km) is at an early stage of development and further research is in progress within JCMM and elsewhere. These results serve as a useful starting point and baseline for testing of new methods and ideas.

8.3.3 Final comments

The results are encouraging. They show quantitatively that improved resolution (down to 1-km gridlength) can have a positive impact on precipitation forecasts, in particular for accumulations over spatial scales larger than ~20km. This is in agreement with the impression obtained from previous subjective analyses of case studies that there is the potential to significantly improve rainfall predictions over the scales of small to medium sized river catchments. They also show that there is still plenty of work to do for that potential to be fully realised.

The scale-selective verification approach using fractions over different sized squares is a valuable tool for assessing the skill of NWP-model precipitation forecasts. It is also a suitable method for determining the most appropriate scales on which to generate output products, especially for the probability maps as they are constructed using the same method as the verification.

The main caveat is that this investigation has only examined a small number of events and that conclusions are being drawn within that limitation. Further cases need to be examined, particularly of severe events that were lacking over the region of interest in the unusual summer of 2003. The summer of 2004 will provide many of those cases.

9 The Boscastle Flood

At the time of writing this report, a serious flash-flood occurred in the village of Boscastle close to the north coast of Cornwall (SW England). The operational 12-km gridlength mesoscale model failed to predict the localised intense rainfall that led to this event. A trial version of a 4-km gridlength model was run on the day, and that did produce much higher rainfall accumulations in the Boscastle area. Since then, further high-resolution forecasts have been run at the Joint Centre for Mesoscale Meteorology (JCMM) at Reading to see if they would have been capable of providing more of a warning.

The purpose of including this section is to show how a model with a 4 or 1-km grid spacing, with the appropriate output diagnostics, could have drawn attention to the possibility of the very high rainfall totals that occurred. There will be little attempt here to discuss the reasons why the thunderstorms formed in the location they did or the particular merits or drawbacks of the individual high-resolution forecasts. Such discussions are for a later date. The intention is to back up previous conclusions with further evidence of how a high-resolution NWP forecast system has the potential to deliver improved forecasts of intense rainfall events.

9.1 The event

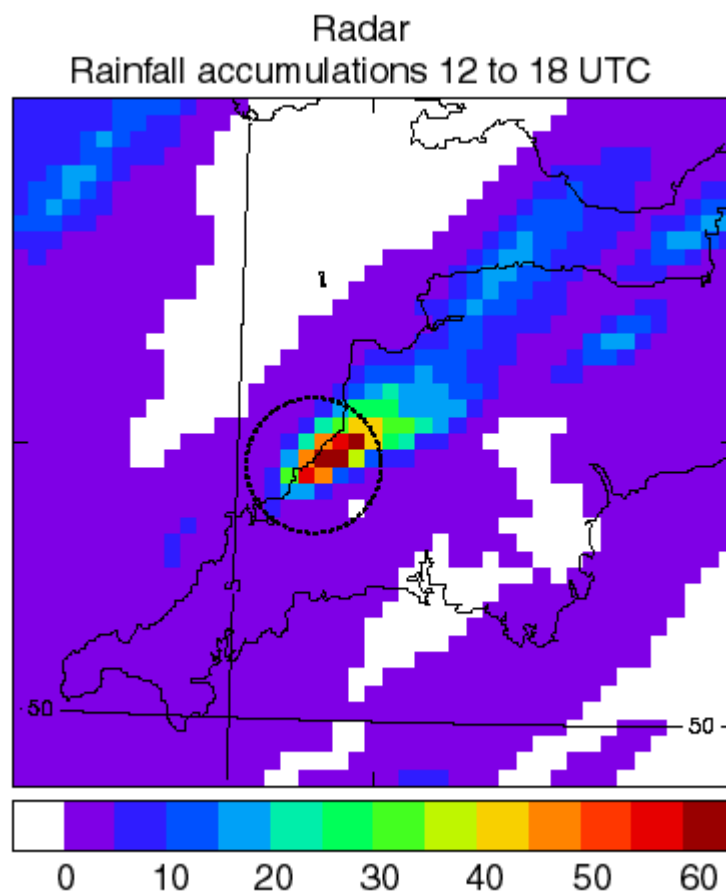


Figure 36. Rainfall accumulations measured by radar over southwest England during the period 12 to 18 UTC 16th August 2004. The radar data were processed by the Nimrod system (ref) and are displayed on 5x5 km pixels.

Intense thunderstorms began over the area around Boscastle at around 10.30 UTC on 16th August 2004 and continued until around 16.30. The showers tended to form in the same location to the southwest of the village and then advect northeast with new cells developing behind. The result was a localised quasi-stationary band of heavy rain that continued for several hours. Pulses of torrential rain fell as new cells formed and moved through. The flooding occurred because the rainfall was concentrated over the small catchment containing the stream that flows through Boscastle.

Figure 36 shows the rainfall accumulations measured by radar from 12 to 18UTC, when most of the rain fell. Accumulations over 5x5km pixels exceeded 70 mm. On a higher-resolution grid with 2x2km pixels, the highest accumulation measured exceeded 128 mm. Rain gauges measured rainfall total in excess of 150 mm. The highest point-total in the period between 11.30 and 16.30 UTC was 181 mm at Lesnewth around 3-4 km east of Boscastle. (this information was supplied to the National Climate Information Centre (NCIC) from the Environment Agency – it has not yet been quality controlled).

9.2 Forecast Products

A selection of diagnostic products from forecasts run with gridlengths of 12, 4 and 1km are now presented. Products are shown instead of the raw model output so that the most important information is displayed on the spatial and temporal scales that may have had reasonable predictability. We expect that a high-resolution or storm-scale model will be able to provide useful information about the amount of rain that will fall somewhere in an area over a period of time, but that the exact size and location of each individual shower is beyond the limit of predictability (see the stage 3 report for more discussion). Suitable diagnostic products should be an integral part of a storm-scale modelling system. The raw output should only be seen by experienced users who understand the predictability issue.

Accumulations over squares

The first product is the maximum rainfall accumulations to occur over 24x24km squares from each of the forecasts and from radar (Figure 37). The accumulation period is 6 hours for the radar and the 12km forecast, but only 3 hours from the 4 and 1-km forecasts. This is because the original 4-km forecast only went out to 15 UTC. It is justifiable because the main period of rain only lasted until 16 UTC so only 1 hour is missed and the heaviest rain predicted by the models occurred around an hour early anyway. It is clear that the 12-km (operational) model failed to predict any rainfall accumulations that were even close to that observed by radar. In fact the 12-km forecast did not even manage any accumulations greater than 10 mm. In contrast, both the 4 and 1-km simulations did predict very high accumulations, and to within the limits of the 24x24km squares, the forecasts were spatially very accurate. The 4-km forecast produced higher totals, that were closer to the radar, than the 1-km forecast, however we should remember that these were only 3-hour forecasts so it does not necessarily mean that the 4-km model was better out to 18 UTC. The point being made here is that both the high-resolution models were capable of producing large rainfall totals in the area of interest, whereas the 12-km model was not.

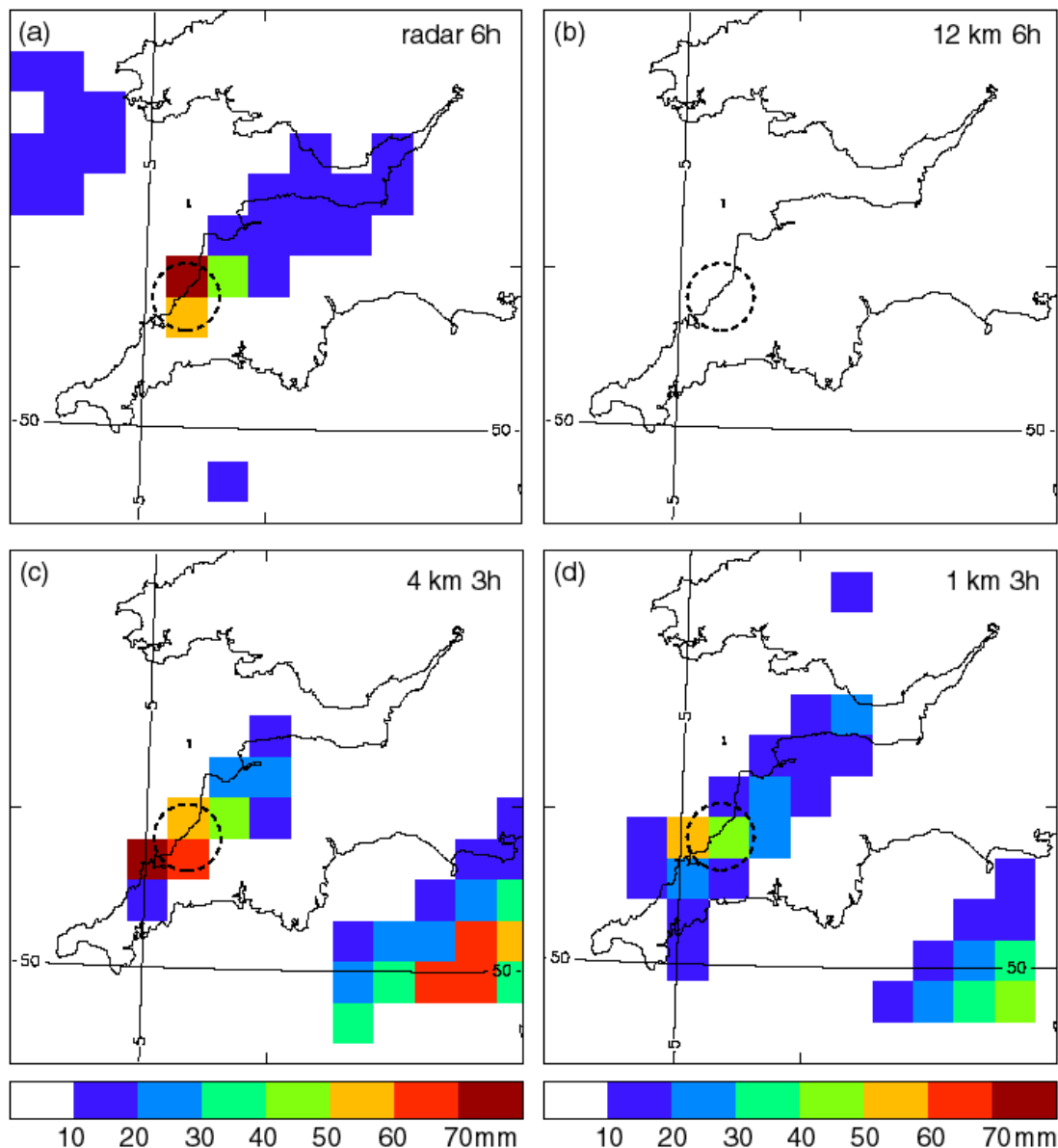


Figure 37. 16th August 2004. Peak accumulations within 24x24 km squares from (a) radar 12 to 18 UTC, (b) 12-km gridlength operational forecast 12 to 18 UTC, (c) 4-km gridlength forecast 12 to 15 UTC, and (d) 1-km gridlength forecast 12 to 15 UTC. All the forecasts started from 00 UTC. The dashed circle is 20-km radius centred at the village of Boscastle.

Following on from the maximum accumulations within a square, the extreme maximum accumulations within a square are shown in Figure 38. They are the highest accumulations that would be possible over a pixel if the heaviest core of rain within a shower or small area persistently fell on the same pixel (section 5.2.1). This product gives the highest totals when heavy showers are slow moving or tend to re-generate in roughly the same place. It is these situations that are most likely to lead to flash flooding, as occurred at Boscastle. The extreme accumulations have only been generated from the 4 and 1-km model forecasts. It made no sense to use this algorithm at 12 km because the grid squares are larger than the individual showers. It has not been applied to the radar data (although it could have been) because the radar is supposed to be the ‘truth’ and we are not attempting to find an extreme realisation of what actually occurred.

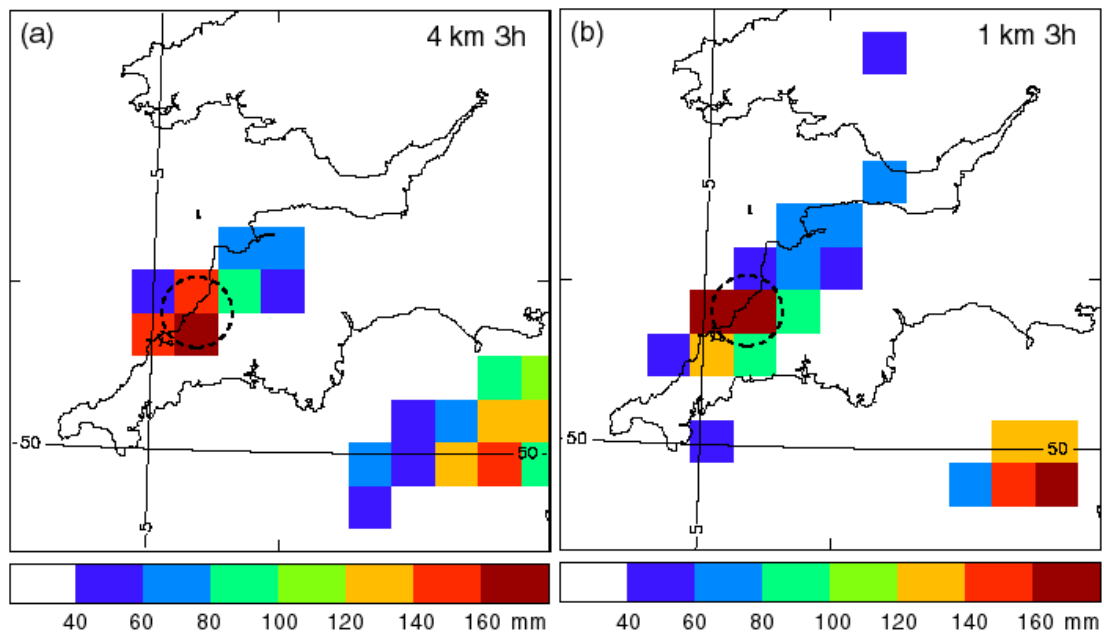


Figure 38. Extreme rainfall accumulations within 24x24 km squares over the period 12 to 15 UTC 16th August 2004 from (a) 4-km gridlength forecast, and (b) 1-km gridlength forecast starting from 00 UTC. The dashed circle is 20-km radius centred at the village of Boscastle.

Both the 4 and 1-km forecasts gave an extreme accumulation of more than 160 mm in the Boscastle area. It is noticeable that the 1-km forecast gave the highest extreme values even though the maximum values were lower (Figure 37). The reason is due to the storms being better resolved at 1 km and therefore more realistic on small scales, so the method had the opportunity to generate locally higher extreme values. Although this product can not be easily verified because it is presenting a scenario that is by definition unlikely to occur, it is interesting that the totals were very similar to the point values measured by rain gauges for this particular extreme event.

Accumulations over river catchments

Another way of presenting the information is to display on river catchments or as in these examples EA warning areas. For the rest of this section the warning areas will be referred to as 'catchments'. This is reasonable for most of the areas displayed as they tend to naturally define river basins. However, we should be aware that the area containing Boscastle does not represent a single catchment, but rather, numerous very small catchments along a coastal strip, of which the local Boscastle catchment is one. Some products have been specifically designed to account for display areas containing several smaller catchments.

All the products shown here have already been discussed in section 5.2.2.

The catchment-average accumulations from radar and the three forecasts are presented in Figure 39. Two features stand out: First, the degree to which the catchments can be represented is determined by the resolution of the model. The 12-km model is unable to represent the catchments properly, the 4-km is much better and the 1-km does the job properly for catchments of this size. Second, the 12-km model does not give any indication of higher average values over the catchments in the vicinity of Boscastle, rather it gives low totals everywhere across the southwest peninsular. The 4 and 1-km forecasts, in contrast, did predict higher average accumulations over catchments in the vicinity of Boscastle, though without indicating a serious risk for the Boscastle area.

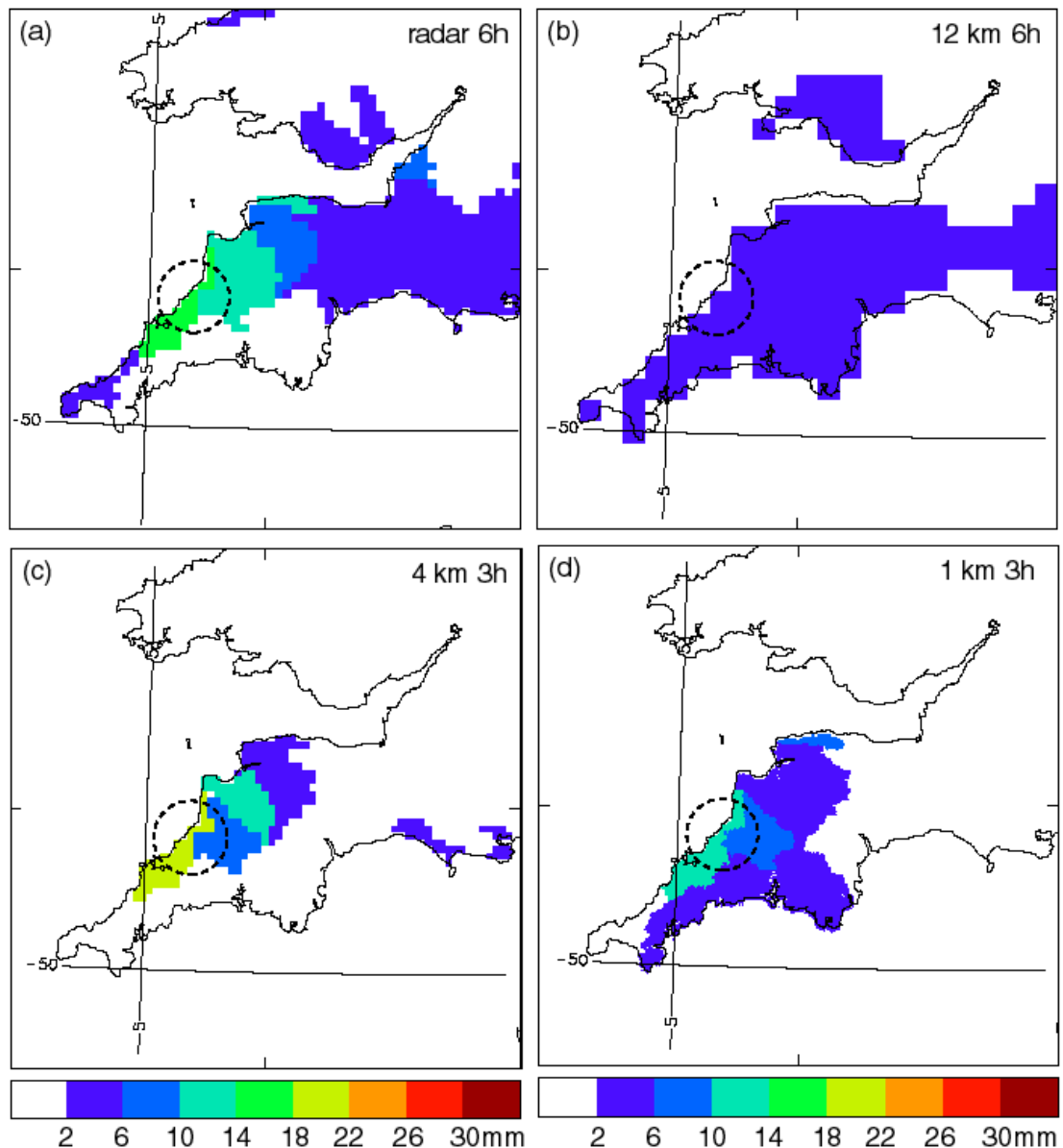


Figure 39. 16th August 2004. Catchment-average accumulations within from (a) radar 12 to 18 UTC, (b) 12-km gridlength operational forecast 12 to 18 UTC, (c) 4-km gridlength forecast 12 to 15 UTC, and (d) 1-km gridlength forecast 12 to 15 UTC. All the forecasts started from 00 UTC. The dashed circle is 20-km radius centred at the village of Boscastle.

The problem is that the Boscastle catchment is much smaller than those displayed in Figure 39. It is a very small part of the large ‘catchment’ that follows the coast (14–18 mm in Figure 39(a)), which is really comprised of several independent smaller catchments. The forecasts of average accumulations over such large areas are not very helpful when there is a concern about more localised events. We should really be presenting information about smaller areas. However, we know that this is dangerous because any forecast is subject to spatial errors and these become more serious over smaller scales. One solution to this problem is to present the model output as shown in Figure 40. The pictures show the highest average rainfall accumulations that are possible over areas smaller than the size of the catchments themselves, given an error in the forecast of up to 12 km. An example of the same product was shown in Figure 15(b).

Now we see that the 4 and 1-km forecasts predicted average accumulations of over 40 mm within smaller areas inside the larger catchments, without being specific about exactly where. The threat of high totals over small catchments is now made evident in a way that does not attempt to give the impression of greater spatial accuracy than we know the forecasts are capable of. Again, we see that the 12-km model was neither capable of representing the catchments properly or indicating a threat of locally high rainfall totals.

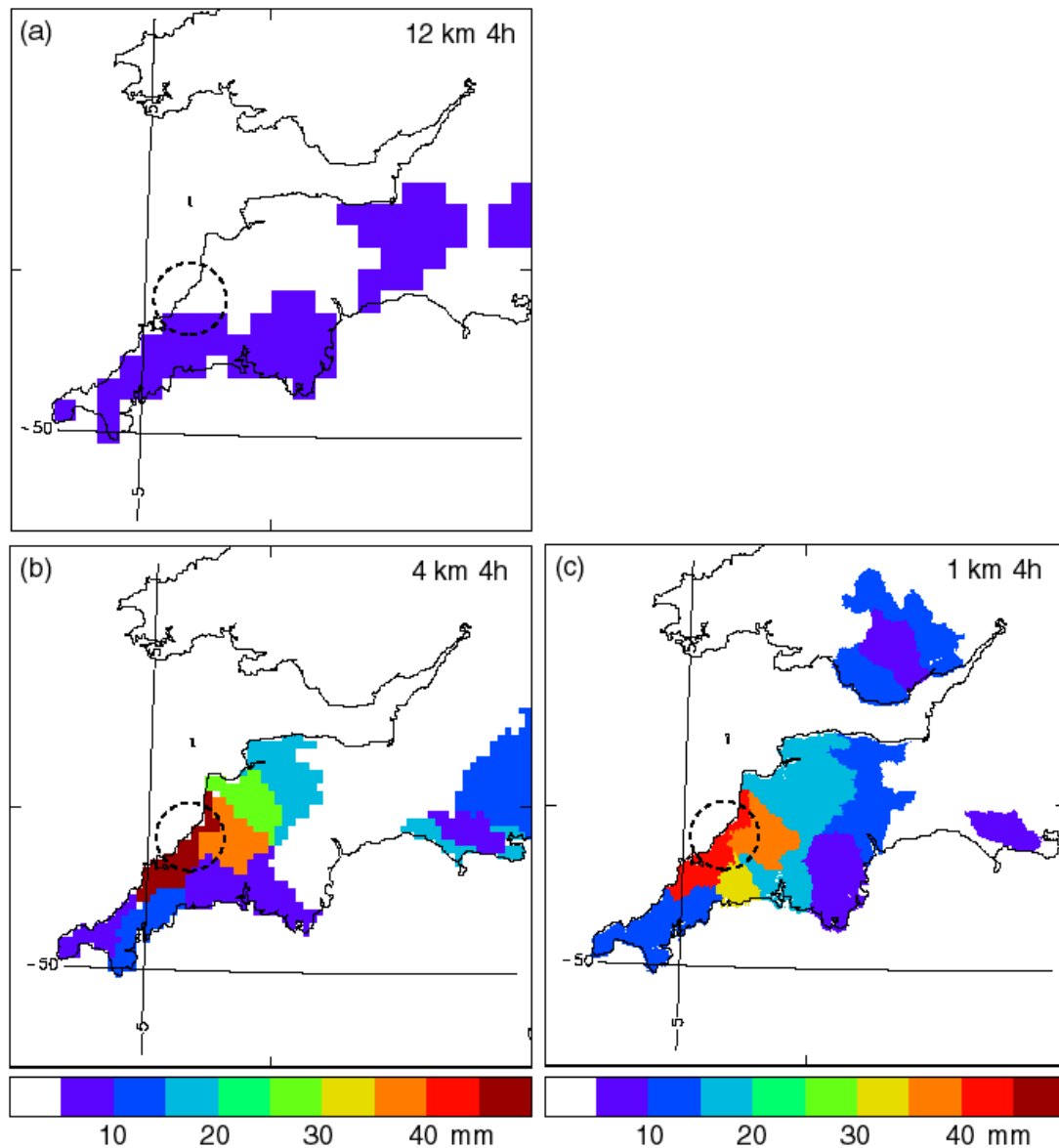


Figure 40. 16th August 2004. Worst-case scenario average accumulations displayed on river catchments for (a) half-catchment size areas from a 12-km gridlength operational forecast 11 to 15 UTC, (b) one fifth catchment size areas from a 4-km gridlength forecast 11 to 15 UTC, and (c) one fifteenth catchment-size areas from a 1-km gridlength forecast 11 to 15 UTC. All the forecasts started from 00 UTC. The dashed circle is 20-km radius centred at the village of Boscastle.

The uncertainty in a forecast can be presented by means of probabilities. Figure 41 shows probabilities of rainfall amounts exceeding 50 mm over the period 12 to 15 UTC from two 4-km forecasts that started at 00 and 03 UTC. The way the probabilities are generated is discussed in section 5.2.3 and more fully in the stage 3 report. The purpose of combining two forecasts is to obtain a more realistic view of the forecast uncertainty. The picture show that, for this particular case, a combination of probabilities from both forecasts

picked out the Boscastle area as being more at risk from very high accumulations than elsewhere. It reveals that there was predictability from forecast to forecast (both forecasts contributed to the non-zero probabilities around Boscastle) for high rainfall totals in that region. This approach of blending the most recent forecast with older forecasts into a ‘time-lag’ ensemble is not new. The advantage we now have with high-resolution models is that probabilities that can be generated from each individual forecast and then combined into a single picture of the uncertainty.

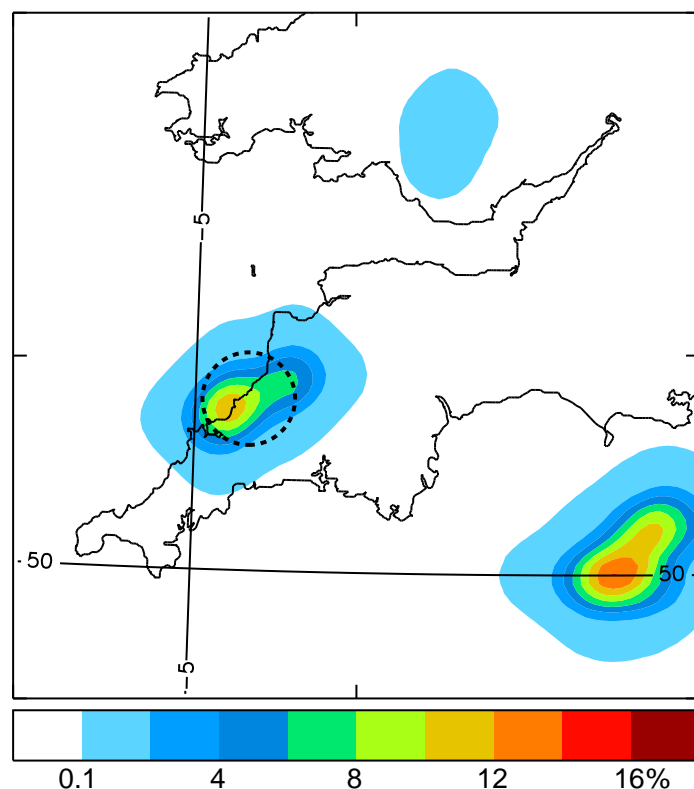


Figure 41. Probabilities of rainfall accumulations exceeding 50 mm over the period 12 to 15 UTC 16th August 2004 generated from two 4-km gridlength forecasts starting at 00 and 03 UTC. The dashed circle is 20-km radius centred at the village of Boscastle.

9.3 General comments

The Boscastle flood is a high profile example of a flash flood caused by rain from localised thunderstorms falling into a small river in a fast response catchment. High-resolution simulations (4 and 1 km grid spacing) were able to indicate that there was a possibility of very high rainfall totals falling in a short space of time over the Boscastle area. The operational 12 km forecast could give no such indication (and would not have been able to even if benefiting from statistical downscaling). Post-processing techniques have been applied to the high-resolution model output to highlight the risk of a severe event without revealing more spatial detail than we can reasonably justify. The use of scenarios and probabilities is an essential part of the interpretation of these forecasts.

Whilst it is true that the high-profile nature of the event and the success of the high-resolution simulations have led to its inclusion in this report, the intention has not been to draw conclusions from this case in isolation. When set in the context of the findings from the previous case studies and verification scores, it serves to back up our perception that high-resolution models can provide more useful warnings of flood-producing rainfall events than current operational models, provided that care is taken in the interpretation of the output.

10 Conclusions

The aim of the project was to investigate the ability of a storm scale configuration of the Met Office NWP model to predict flood-producing rainfall up to 12 hours ahead and to develop appropriate tools for interpreting and presenting the predictions so that they enhance operational flood prediction capabilities. The work was carried out in a research framework. It is not be possible to run such a model operationally at present, but may be within a few years. This section will be used to give an overview of some of the achievements and key results that address the objectives of the project, then discuss recommendations for future work and give a brief summary of the most important outcomes.

The conclusions are split into four parts.

- (1) Achievements
- (2) Results
- (3) Recommendations
- (4) Summary

10.1 Achievements

1. The first and most basic achievement was to run the model successfully and produce realistic simulations. Several case studies were simulated using a grid spacing of 4, 2 and 1 km. At the start of the project, we were stepping out into the unknown. The new non-hydrostatic version of the model, necessary for the high-resolution simulations, had not yet even become operational at global or mesoscale resolution (grid spacing of ~ 60 and 12 km). It should be acknowledged that without the considerable effort that went into the development of the new model (non-hydrostatic and semi-lagrangian dynamics) within the Met Office in the years running up to this project it could never have begun.
2. Once it was established that the model worked, the sensitivity of forecast performance to varying key tuneable parameters was examined. The four initial case studies were used as a test bed for these experiments. Systematic testing was also performed within the High Resolution Trial Model project to allow informed decisions to be made about how the model should be configured. See section 3 for more information.
3. One of the problems that was anticipated and did indeed turn out to be a significant issue was how convection should be represented when it is sometimes only partly resolved on a high-resolution model grid. Results from the case studies showed that, for the 4-km model in particular, the use of the convective parametrization scheme could cause serious problems and that simply switching off the convection scheme was not the answer either. A workable solution was obtained means of a modification that restricted the operation of the convection parametrization. See the stage 2 interim report.
4. Diagnostic products were developed to optimise the usefulness of storm-scale model forecasts for flood prediction. They have been designed to extract the most useful and reliable information from high-resolution model rainfall forecasts, and can then be used to aid manual interpretation or be incorporated into an automated system or be used as input to hydrological models. See the stage 3 interim report.
5. A methodology has been developed to allow precipitation forecasts to be verified against radar over different spatial scales. Such an approach is vital for assessing the performance of a storm-scale model if we wish to determine the scales over which a model is sufficiently reliable for flood prediction. Appropriate products can then be produced. Traditional grid point by grid point verification methods

can only verify scales that we already know are unreliable. See the stage 4 and 5 interim reports.

10.2 Results

The storm-scale model produced more realistic forecasts

- A high-resolution NWP model (grid spacing 4km or less) is capable of producing much more realistic simulations of convective rainfall events than the current operational mesoscale model (grid spacing 12 km). The ‘storm scale’ model forecasts (1-km grid spacing) gave the most realistic representation of rainfall patterns and intensity. The evidence for this conclusion comes primarily from the visual examination of the 9 events presented in the project reports, and also from inspection of further cases run within the High Resolution Trial Model (HRTM) project. This result was anticipated because a storm scale model by design is able to resolve features that can only be represented by a convection scheme in a 12-km grid-length model. Even so, the results are impressive.

The storm-scale model produced more accurate forecasts on scales applicable for flood prediction

- A new verification methodology for comparing rainfall accumulation forecasts with radar in a scale-selective way has been used to assess forecast performance over a number of events. The results show that the forecasts performed better at all resolutions when the target area was larger, i.e. it was easier to predict that it would rain somewhere within a large area than a small one. Most important though, was the discovery that the high-resolution (1 and 4km) forecasts out-performed the 12-km forecasts over nearly all spatial scales. In particular, they gave more accurate predictions of higher rainfall accumulations over areas the size of small to medium sized river catchments. The best scores for these criteria were obtained from the 1-km grid-length model. Since this project is concerned with the use of a storm-scale model for flood forecasting, these are encouraging results. Not only can a high-resolution model produce more realistic simulations, but it is also more accurate, at least up to 7 hours ahead, over scales that are important for flood prediction.

Diagnostic products are vital for forecast interpretation

- The use of diagnostic products is essential for a meaningful presentation of rainfall forecasts from a storm-scale model, especially when used in a flood forecasting context. The objective verification results showed that over smaller spatial scales, even the high-resolution simulations can sometimes have little skill. Certainly, any scale close to the grid-scale of the model can not be relied upon. Sensitivity studies have revealed that small changes to the values of some parameters will have an impact on rainfall patterns in a forecast. The problem in interpretation arises because high-resolution forecasts can look so realistic and it is tempting to believe the detail, when in fact a different forecast outcome could look just as realistic. Diagnostic products that were designed to exploit the advantages of the storm-scale model for the purpose of rainfall prediction, but still take forecast uncertainty into account, have been shown to add considerable benefit to the raw model output. Any storm scale forecast model should have an appropriate post-processing system built in, otherwise the full potential of the model as a forecasting tool can not be realised and there is a danger of misinterpretation. There is considerable scope for products to be developed that can generate different forecast scenarios for input into hydrological models.

Data assimilation at high resolution requires further development to give consistent benefit

- Scale-selective verification was used to examine the impact of data assimilation. The results showed that data assimilation at high resolution had a better fit to radar at the early stages of the forecasts, but tended to lead to a poorer distribution of rainfall after a few hours compared to forecasts that had to ‘spin up’ from the 12-km grid-length model fields. On the other hand, the spin-up forecasts produced too much rain after a few hours and this was improved by the data assimilation. The 1-km forecasts with data assimilation scored higher than the other forecasts for a rainfall accumulation threshold of 8 mm (the highest accumulation it was possible to examine over these events). These are very preliminary results, though nevertheless a useful measure of current capability. The data assimilation used at 4 km was essentially the same as that used operationally at 12 km (3DVAR and MOPS cloud analysis and latent heat nudging, see interim report 5). The data assimilation at 1 km involved the addition of increments from 3DVAR at 4 km. More appropriate methods for high-resolution models are being developed at the Joint Centre for Mesoscale Meteorology (JCMM). It is clear that some kind of data assimilation on the high-resolution model grid is essential if we are intending to use such a model for very short range forecasting up to a few hours ahead. It is not clear whether the same can be said for forecasts beyond a few hours. Care should be taken with these findings however, they have been obtained from a small sample and further objective testing is required.

10.3 Discussion and recommendations

The results have shown that a storm-scale model (grid spacing ~1 km) does have the potential to deliver more accurate predictions of flood-producing rainfall events than our current operational systems. For that very important reason, it is worth continuing the development of such a model towards operational implementation.

The advantage of a very high-resolution model is that it can simulate the physics and dynamics of individual storms in a way that is just not possible in a coarser-resolution model, which relies largely on a convection scheme to represent showers. The case studies showed that the 12-km gridlength operational mesoscale model is much less able to simulate local flood-producing thunderstorms than a 1-km gridlength model, and even if statistical downscaling were to be applied to the mesoscale model output it would still not be as successful because the distribution of the rainfall is poorer on the scales of interest. This implies that a storm-scale model is more suited to the prediction of flood-producing storms than even an ensemble of mesoscale model forecasts, which would simply not be able to predict some events whatever the ensemble size.

However, an ensemble approach is still worth investigating. We know that there is uncertainty associated with rainfall forecasts whatever the resolution of the model. Some of the errors in a high-resolution model come from information that is passed through the domain boundaries from a coarser-resolution model. Disturbances in the flow on the scale of a few tens to a few hundreds of kilometres can have a significant impact on the distribution of convection. These are situations in which the development of a coarser-resolution ensemble could provide alternative scenarios for larger-scale dynamical forcing (e.g. frontal zones) and a limited number of storm-scale forecasts could be run from a small sample of the ensemble members. However, that kind of approach would demand considerable computer resources and hence be impractical as an operational system for

the foreseeable future. The additional use of an ensemble of storm-scale simulations would be necessary if we wish to represent the unpredictable nature of smaller scales and that would have to be an even longer term project.

Until a high-resolution ensemble system is a viable option we will have to rely on post-processing individual forecasts to represent the forecast uncertainty across a variety of scales. The products developed within this project are effective tools for doing that and for displaying the most relevant information needed for flood prediction and input into hydrological models. It is essential that the generation of appropriate rainfall products becomes an integral part of any future high-resolution model forecast system if we wish to avoid the misinterpretation of precipitation forecasts. Further work is needed in this area. The products that have already been developed need to be presented to potential users for comment and made available as output from a trial system to establish how well they work in practise. At present we do not know what scales to generate the products over in order to obtain the best trade off between usefulness and accuracy. The examples shown in the reports from this project have been produced using an educated guess at the most appropriate sampling areas. Scale-selective verification needs to be performed over a large number of events to determine those scales over which a high-resolution model has sufficient skill to be useful, so that products can be generated accordingly. A system is required that will feed verification results into the product generation process and in turn verify the skilfulness of those products. Emphasis should be placed on the use of rainfall products that feed different forecast scenarios as input into hydrological models. Examples of such products have been presented in this report. This should be seen as an important function of a storm-scale model for flood-warning. Collaborative work will be needed to integrate high-resolution rainfall predictions (from both numerical models and nowcasting systems) into hydrological systems.

Data assimilation at high-resolution is an area of considerable difficulty and ongoing research. It must remain a top research priority. Results have shown that it is simply not satisfactory to leave out data assimilation in a high resolution model if it is to be used for short range rainfall predictions (up to 6 hours ahead) because it will not be capable of representing convective showers correctly over the first few hours. On the other hand, it appears that the data assimilation techniques tested so far have had a tendency to make the spatial accuracy of rainfall forecasts worse after a few hours by making the dynamical fields less coherent. This despite a better fit to radar at the start. Part of the problem comes about for the very same reason we need diagnostic products to present high-resolution rainfall forecasts. The errors in a high-resolution precipitation forecast grow more rapidly at small scales than in a coarser resolution model, which is incapable of resolving those scales. I.e. a storm-scale model is able to generate small showers, but is very unlikely to have them in exactly the right place, and the positional error gets worse with time. Data assimilation techniques rely on a previous forecast to give a 'best guess' before new information is added. If that previous forecast is long compared to the time it takes for small-scale errors to grow, then it may contain a considerable amount of spurious information. The standard continuous assimilation approach, in which this process is repeated again and again, will compound the problem and grow the errors to larger scales. Given current data assimilation methods, there are two ways of reducing this problem; one is to have a much more rapid update cycle – i.e. use a much shorter forecast so the errors have had less time to grow; the other is to start from coarser-resolution fields when performing the high-resolution data assimilation. In the context of a storm-scale model, it means examining two options; either starting a new forecast with data assimilation every hour or even 15-30 minutes instead of every 3 hours, or applying data assimilation on the 1-km grid but using a forecast from the 4-km model as the starting point (which in turn uses a forecast from the 12-km model to start with). The first option is computationally expensive and that could be an obstacle, although improvements in data assimilation methods may mean that a longer update time will become feasible. The second option might appear less attractive because high-resolution information is not being passed from forecast to forecast, but it does mean that a longer update cycle can be justified. New data

assimilation methods, which are designed to spread information in time as well as space, may make this discussion less relevant, but they too are computationally expensive.

The discussion, so far, has concentrated on the use of a storm-scale model as a tool on its own. That would probably not be the best way forward. The current operational approach for short-range precipitation forecasting is to use the Nimrod or Gandolf nowcasting systems for the first few hours and then blend in information from the 12-km mesoscale model after that. The rationale is that a nowcasting system is more accurate at the start (as well as being available earlier), and the numerical model becomes more skilful later. There is no reason to suppose that the situation should be any different for a storm-scale model given the results from the data assimilation experiments. We still expect advection nowcasting to be more accurate at the beginning of a forecast. For how long it will be better, we don't know. The hope is, of course, that the combination of high-density observation (e.g. radar), improved data assimilation methods and a rapid update cycle will eventually deliver a storm-scale modelling system that out-performs a nowcasting system at all times. Until that is the case (if ever) we should use both. The work on storm-scale modelling is timely because it coincides with the development of the stochastic nowcasting system (STEPS). The STEPS system is designed to run an ensemble of short-range forecasts (2-km grid spacing) to account for forecast uncertainty and can therefore generate a probability distribution of predicted rainfall. Products from a storm-scale model can also be used to generate rainfall probability distributions and convey information about forecast uncertainty. Research is needed to investigate how STEPS and a storm scale modelling system can be blended seamlessly to provide the most useful output for flood warning.

The development of a technique for verifying precipitation forecasts over different spatial scales has been successful. It has been used to assess precipitation forecasts and analyses from the operational mesoscale model (12-km grid spacing) over an entire year (2003) and produced useful results. It was needed because the alternative of verifying at the grid scale is not sensible and is particularly meaningless for a 1-km model in which we expect the small scales to quickly become unpredictable, but the larger scales to retain some skill. The results have given a helpful insight into model behaviour. Further work is needed to expand the verification approach and compare with other techniques (e.g. Casati et al 2004). Work is also required to address the issue of uncertainty in the radar. Most importantly, verification of a larger sample of high-resolution forecasts is necessary. The benefit of a larger sample has been highlighted by the results from the mesoscale model verification.

The case-study simulations have revealed aspects of high-resolution model behaviour that require further attention. The most important of these is the way we represent the convective clouds that can not be resolved on the grid. The restriction on the activity of the convection scheme introduced into the 4-km gridlength model has had beneficial results, but there are still problems. Resolved showers initiate too late and become too large and intense. Some kind of representation of convective turbulence due to unresolved cumulus clouds is necessary at 4 km and to a lesser extent at 1 km. Research within the Met Office at JCMM is aiming to tackle this problem. In the meantime, there are ways in which the restriction on the convection scheme can be improved and these should be examined. The introduction of stochastic noise into the model as a way of bringing forward shower initiation should also be investigated.

Another fundamental problem is the time it takes for showers to develop in the flow coming in through the edge of the domain. This can often lead to a strip close to the edge where there is no convective rain. If the flow into the domain is strong, the strip can be wide and seriously affect a forecast. The problem is worse in a 4-km gridlength model because the 12-km model that supplies the boundary information hardly resolves any convection. However, it may appear worse in a 1-km model because the domain has to be smaller. The development of a variable resolution model is underway at JCMM, and that

may ultimately help considerably because there will be a much more gentle change in resolution rather than a sharp boundary. With the current set up, the introduction of small-scale noise in the boundary region is an option to be tested. The most effective way of dealing with the problem may be to use a larger domain and only present output from a smaller inner area. That will make a model more expensive to run. Systematic testing of the impact of domain size is needed before decisions can be made. It is quite likely that the impact of the boundaries will extend throughout the domain, even where showers are able to develop.

The sensitivity of 1-km model rainfall forecasts to changes in the number of vertical levels, the amount of diffusion, the introduction of targeted diffusion, changes to the cloud microphysics, the representation of soil moisture, the length of the time-step and the frequency of boundary updating are further issues that will need re-visiting as development continues.

The case studies have provided examples of successful high-resolution forecasts of flood-producing situations when the 12-km model was poor. However, we do not really know how sensitive a high-resolution forecast of a severe event is to small changes in model parameters or initial conditions, and therefore how much it can be relied on. We know that small scales are inherently less predictable, but do not know what scales are predictable for a given event. For example, in the case of the localised storm over East Anglia (case 4, section 4.5), the prediction of a storm was critically dependant on the number of vertical levels. Does this mean that it was an inherently unpredictable event or that there was a problem with the model when using more levels? We might expect the Boscastle flood to be more predictable because it appears to have been tied to the local orography, but that has yet to be shown. There is a need for detailed investigation of high-resolution simulations of severe events to determine what it is that the model has to get right to produce good forecast, where the model has deficiencies, and how sensitive the forecast is to changes in initial conditions or model formulation.

11 Summary

The objective of the storm-scale modelling project was to investigate the ability of a storm scale configuration of the Met Office NWP model (grid spacing down to ~1 km) to predict flood-producing rainfall up to 12 hours ahead and to develop appropriate tools for interpreting and presenting the predictions for the benefit of operational flood prediction. A number of case studies were chosen to examine high-resolution model simulations of a variety of convective events.

The results have provided evidence that a storm-scale model does indeed have the potential to deliver a significant improvement in our ability to predict high-impact convective rainfall events. Even when starting from 12-km gridlength fields, the 1-km gridlength simulations were able to represent local severe storms much better than the operational 12-km model.

Care must be taken in the way output from a storm-scale model is interpreted. The detail can be misleading if taken at face value because of the unpredictable nature of small scales - it is fundamentally impossible to predict the exact locations of all individual showers (except perhaps over an extremely short forecast period) even when a forecast is good. For that reason, diagnostic products have been developed to present the rainfall output on scales that are more predictable. The products are designed to be appropriate for use in an operational flood-warning environment and can also provide input to hydrological models.

A new verification method has been developed to determine the skill of precipitation forecasts on a variety of scales. Such an approach is essential if rainfall forecasts are to be assessed properly and it provides a means of determining the scales over which to generate reliable products for customers.

The value of assimilating observations into high-resolution models (1 and 4-km grid spacing) using current operational methods has been examined. The results showed that data assimilation gave a significant improvement over the first few hours of the 1 and 4-km forecasts, but the improvement was not maintained beyond that time. The findings suggest that high-resolution data assimilation is essential for short-range precipitation forecasts, but further research is needed to find techniques that will lead to better longer-range predictions.

The main recommendation of this project is that the development of a storm-scale model should continue and move towards operational implementation. The potential exists for a significant improvement in flood-warning capability. However, considerable research is still needed to fulfil this potential. The main focus of effort is needed in high-resolution data assimilation. Further work is also required in the representation of sub-grid convection, cloud microphysics, the development and use of output products for flood-prediction, the use of scale-selective verification methods, studies of predictability and techniques for blending precipitation output with a stochastic nowcasting approach.

12 Acknowledgements

I would like to thank my project board Brian Golding, Roderick Smith and Peter Clark for their valuable input. I would also like to thank Humphrey Lean and Andrew Macallan for model suite development at JCMM, without which this project would have been considerably more difficult. Discussions with Mark Dixon and Richard Forbes (JCMM) about data assimilation and cloud microphysics were very helpful as were discussions with Clive Pierce (JCHMR) about nowcasting methods. A very big thanks goes to Peter Panaji (JCMM) for retrieving the bulk of the operational mesoscale model required for verification in section 7. The project was half funded by Defra.

13 References

13.1 Project stage reports.

Stage 1 report:

Roberts, N. M., 2003: Results from high-resolution simulations of convective events. *JCMM internal report.*, 140. (*NWP Technical Report.*, 402)

Stage 2 report:

Roberts, N. M., 2003: The impact of a change to the use of the convection scheme in high-resolution simulations of convective events. *JCMM internal report.*, 142. (*NWP Technical Report.*, 407)

Stage 3 report:

Roberts, N. M., 2003: Precipitation diagnostics for a high resolution forecasting system. *JCMM internal report.*, 143. (*NWP Technical Report.*, 423)

Stage 4 report:

Roberts, N. M., 2004: Measuring the fit of rainfall analyses and forecasts to radar: Strategy for stage 5 evaluation. *JCMM internal report.*, 146. (*NWP Technical Report.*, 432)

Stage 5 report:

Roberts, N. M., 2004: Verification of the fit of rainfall analyses and forecasts to radar. *JCMM internal report.*, 148. (*NWP Technical Report.*, 442)

13.2 Remaining references

Casati, B., Ross, G., Stephenson, D. B., 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts., *Meteorol. Appl.* 11, pp 141-154.

CEH Wallingford, 2001. PDM Rainfall-Runoff Model. (Incorporates: Practical User Guide, Guide, Practical User Guide to the Calibration Shell, User Manual, Training Exercises), Version 2.1, September 2001.

Cox P., Best M., Betts R. and Essery R., 2001: Improved representation of land surface patchiness in the mesoscale model. *NWP Gazette* March 2001. Available on request or from Met Office web page (at time of writing)
http://www.metoffice.gov.uk/research/nwp/publications/nwp_gazette/mar01/improved.html

Cullen, M.J.P., Davies, T., Mawson, M.H., James, J.A., Coulter, S.C. and Malcolm A., 1997, An overview of Numerical Methods for the Next Generation UK NWP and Climate Model" Numerical Methods in Atmospheric and Ocean Modelling. The Andre J.Robert memorial volume. Edited by Charles A Lin, Rene Laprise and Harold Ritchie 425-444

Forbes, R. M. and Clark, P. A., 2003: Sensitivity of extratropical cyclone mesoscale structure to the parametrization of ice microphysical processes., *Q.J.R.Meteorol.Soc.*, **129**, 1123-1148.

Forbes, R. M., and C. E. Halliwell (2003). Assessment of the performance of an enhanced microphysics parametrization scheme in the Unified Model at 1km resolution. Met Office Milestone report for Research Activity RC12CR. Available on request.

Golding, B.W. 1998: Nimrod: A system for generating automated very short range forecasts., *Meteorol. Appl.* 5, pp1-16

Gregory, D. and Rowntree, P. R., 1990: A mass flux scheme with representation of cloud ensemble characteristics and stability-dependent closure. *Mon. Wea. Rev.*, **118**, 1483-1506

Hand, W.H. 2002: The Met Office Convection Diagnosis Scheme., *Meteorol. Appl.* 9, pp69-83

Jones, C.D. and Macpherson, B. 1997: A Latent Heat Nudging scheme for the Assimilation of Precipitation Data into an operational Mesoscale Model. *Meteorol. Appl.*, 4, 269-277.

Lean, H. W., 2003: High Resolution Trial Model project, results following the completion of stage 3. Available on request.

Lean, H. W., and Clark, P.A. 2003: The effects of changing resolution on mesoscale modeling of line convection and slantwise convection. *Quart. J. Roy. Met. Soc.*, **129**, 2255-2278.

Lorenc, A. C., S. P. Ballard, R. S. Bell, N. B. Ingleby, P. L. F. Andrews, D. M. Barker, J. R. Bray, A. M. Clayton, T. Dalby, D. Li, T. J. Payne and F. W. Saunders. 2000: The Met. Office Global 3-Dimensional Variational Data Assimilation Scheme. *Quart. J. Roy. Met. Soc.*, **126**, 2991-3012.

Macpherson, B. 1996: The impact of MOPS moisture data in the UK Meteorological Office Mesoscale Data Assimilation Scheme., *Mon Wea Rev*, **124**., 1746-1766

Moore, R. J., 1985. The probability-distributed principle and runoff production at point and basin scales. *Hydrological Sciences Journal*, **30**(2), 273-297.

Moore, R. J., 1999. Real-time flood forecasting systems: Perspectives and prospects. In: R. Casale and C. Margottini (eds.), *Floods and landslides: Integrated Risk Assessment*, Chapter 11, 147-189, Springer.

Persson, P. O. G., and Warner, T. T. 1995: The nonlinear evolution of idealized, unforced, conditional symmetric instability: a numerical study., *J. Atmos. Sci.*, **52**, 3449-3474.

Pierce C, Bowler N, Seed A, Jones A, Jones D and Moore R., 2004: 'Use of a stochastic precipitation nowcast scheme for fluvial flood warning and prediction'. Success stories in radar hydrology, 6th international conference on Hydrological Applications of Weather Radar, Melbourne, Australia, 2-4 February 2004. Published by Bureau of Meteorology.

Smith, R.N.B., 1990. A scheme for predicting layer clouds and their water content in a general circulation model. *Q.J.R. Meteorol. Soc.*, **116**, 435-460

Wilson, D. R., Ballard, S. P., 1999. A microphysically based precipitation scheme for the UK Meteorological Office Unified Model. *Quart. J. Roy. Meteorol. Soc.*, 125, 1607-1636.

Zhang, F., Snyder, S. and Rotunno, R., 2003: Effects of Moist Convection on Mesoscale Predictability., *J. Atmos. Sci.*, 60, 1173-1185