Understanding the national performance of flood forecasting models to guide incident management and investment

Steven Wells[a,1], Alice Robson[1], Robert J. Moore[a,1], Steven J. Cole[1] and Alison Rudd[1]

[1]*Centre for Ecology & Hydrology, Wallingford, Oxfordshire, OX10 8BB, UK*

Abstract. The preparation of routine flood guidance statements and formulation of incident management strategies requires national operating agencies to have a firm understanding of the performance of flood forecasting models. Studies of flood forecasting model performance are commonly evaluated on a grouped-catchment or local basis and can lack the analytical consistency required for integration into coherent national assessments. Here, the first nationally consistent analysis of flood forecasting model performance across England and Wales is presented. Application of the assessment framework, accounting for regional and model-type differences, yields a national overview of relative forecasting capability for models in current operational use. To achieve extensive site coverage, information from many existing local performance studies are pooled into a single structure for analysis under a national framework. The performance information spanning a variety of local models is also compared against the area-wide national G2G (Grid-to-Grid) distributed model. An integrated national assessment gives an evidence base of model performance useful for guiding strategic planning and investment in flood forecasting models. A concise single-page Performance Summary has been created for each site model that contains performance statistics, forecast hydrographs and catchment properties to aid operational use. A prototype web portal has been developed to make information on forecasting model performance more accessible and understandable for end-users.

1 Introduction

Understanding the performance of flood forecasting models is essential for their informed use for flood guidance. This paper provides an overview of the first nationwide analysis of flood forecasting model performance across local centre implementations of the National Flood Forecasting System (NFFS) [1], using a nationally consistent assessment framework [2, 3]. The analysis incorporates both local models used within river network models across a region and the Grid-to-Grid (G2G) model [4, 5], implemented within NFFS as an area-wide national model across England and Wales [6]. G2G as a distributed rainfall-runoff and routing model is thus compared against the local models which span across a variety of model types: rainfall-runoff of conceptual and transfer function form, and channel flow routing models of hydrological and hydrodynamic form.

Raw datasets – river flow observations, historical simulations and flow forecasts – were gathered from previous local performance studies and standardised under a single assessment framework.

The study aimed to construct a template for site-by-site performance reporting, containing a variety of different performance measures and graphical displays to be used by staff in both an operational setting or during strategic planning. Collating these results then served as the basis for a national analysis and summary of current model performance. This constituted an extensive national evidence-base of model performance, stratified by model-type, model-group, geographical region and lead-time.

2 Forecast model dataset coverage

The assessment considered local models - used in a network of models representing rivers across a region - and the national G2G model, a distributed grid-based hydrological model with area-wide coverage used to forecast river flows on a 1km grid across England & Wales. The local models comprise of rainfall-runoff models of conceptual (PDM, MCRM, TCM) and transfer function (PRTF) type and channel flow routing models of both hydrological (KW) and hydrodynamic (ISIS, MIKE 11) type. The acronyms used here stand for Probability Distributed Model, Midlands Catchment Runoff Model, Thames Catchment Model, Physically Realisable Transfer Function and extended Kinematic Wave model: for further information see [7] and [8].

[a] Corresponding author: stewells@ceh.ac.uk, rm@ceh.ac.uk

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Output from these models can be in the form of historical simulations, or forecasts that emulate real-time use. In simulation mode, model inputs are transformed to outputs without reference to observed flows. Forecast outputs invoke data assimilation of river flow through ARMA (AutoRegressive Moving Average) error prediction, state updating and direct flow insertion, sometimes in combination. The majority of models employ a 15 minute time-step; some local models for the rivers Severn and Trent (Midlands) and in the South West operate at an hourly time-step.

Liaising with the Environment Agency and Natural Resources Wales, flood forecasting locations with model data suitable for analysis from past performance studies were identified across England and Wales. Figure 1 illustrates the extent of the locations that were included within the study. The manner in which the performance analysis is constructed allows for local models at these sites to be updated, or new models added, making future reporting of performance an efficient process. Figure 1 also includes the spatial distribution of gauged sites at which G2G is assessed.

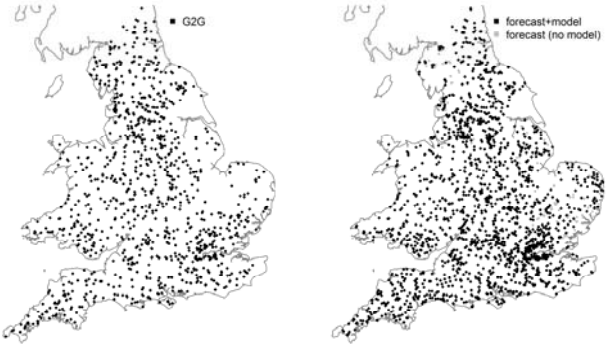

Figure 1. Sites with datasets available for performance analysis of G2G (left) and local models (right, indicating where forecast and model-simulation data are present).

Datasets for 921 local site models are suitable for analysis encompassing 8 model-types and all 8 geographical areas across England & Wales. A further 110 local site models with data were omitted from the analysis for a variety of reasons including site decommissioning, out-of-date model, strong tidal influence, and data too poor to justify analysis. Outputs from G2G have been produced for 1036 gauged sites across England and Wales and this coverage can be compared to the local model locations in Figure 1. A total of 829 are suitable for analysis, with 207 previously judged unsuitable on account of artificial influences affecting the hydrograph or problems with flow gauging. It is noted that G2G has the capability to provide flows everywhere across its model domain on a 1km grid so, in practice, forecasts are available for any ungauged or gauged location. A total of 498 sites analysed had multiple models providing forecasts for a given gauged location, allowing for a direct inter-comparison of model performance.

In collating and analysing the local model datasets, model-related aspects such as rainfall input (observed or forecast), output variable (flow and/or level) and forecast time-step reveal strong regional differences that must be considered when interpreting performance analyses on a national-scale.

The types of rainfall used as input to the models could be:

- *perfect rainfall (*rainfall observations: assumes perfect foreknowledge of future rainfall),
- *forecast rainfall* (uses the archived weather model rainfall forecasts that would have been available at the time) and
- *no rainfall* (zero rainfall assumed).

The most appropriate input type for making a comparison of model performance is *perfect rainfall*. Whilst use of forecast rain mimics the operational setup, it serves to confound the errors due to the model and forecast rainfall. Since model error is the priority for assessment, datasets that use perfect rainfall have been chosen wherever possible. This accounted for 90% of the local models analysed; the remainder employed forecast rainfall. All the G2G modelled flows were obtained with perfect rainfall as input.

For a given river gauging station, observed and modelled data were available as flow or level, and for some models both were present. Where possible, level data were converted to flow via rating equations. Hydrodynamic models, such as ISIS and MIKE 11, operate in both level and flow whereas hydrological models derived from mass-balance equations are flow-based. The majority of local site models analysed had data as level. Model simulation data are only used when there is a matching observed series; in some cases no simulation data were available. The G2G data, both observed and modelled, are in terms of the flow variable.

3 Skill score assessment

The statistical measures employed in the assessment of model performance have been chosen to particularly focus on those relevant to the operational user. An *event* is first defined as the upward crossing of a flow/level threshold. Skill scores can then be constructed to assess the success of a flood model forecasting such events. Additional measures judge the efficiency of a model forecasting the magnitude and timing of a flood peak relative to what is observed.

3.1 Thresholds

In order to define whether a particular flow/level can be classed as an event, a threshold must be defined. The notation Q(T) refers to a river flow Q of return period T years. The special case of Q(2), denoting the median annual maximum flood, is called QMED. For natural rivers QMED has the practical interpretation of aligning to the bankfull discharge at which a river starts to overtop its banks. Where possible, model performance is evaluated at two nationally consistent thresholds: G2G QMED and G2G QMED/2. Here G2G

QMED is estimated from G2G river flows over a period of five water-years.

Local model assessment was undertaken in levels if available. Where the G2G thresholds could not be converted to levels, or there were too few crossings of the G2G level thresholds by the local model, the observed flows from the local model datasets were used to derive the thresholds. This was implemented using a Peak-Over-Threshold analysis with a threshold progressively decreasing from the maximum to minimum observed flow, until at least five (ten) peaks were identified for the upper (lower) threshold.

3.2 Magnitude tolerances and lead-time windows

When using threshold crossings to define events, it can be useful to be less stringent and allow a tolerance in magnitude around the threshold value and/or a lead-time window.

If a magnitude tolerance is applied, an event is defined as an upward crossing anywhere within a given magnitude range. Magnitude tolerances were set on the thresholds at ±20% for flow, and either converted to an equivalent level value using rating equations, or if not available set to ±0.2m.

Model and observed threshold crossings within an allowed lead-time window $L \pm \Delta t$ of a target lead-time L are compared to produce skill scores [9, 10] to assess flood forecasting model performance. The window width allowance Δt is chosen to increase with lead-time from ±1 hour at 1 hour lead-time to ±4 hours at 36 hours lead-time: see [2, 3] for details.

3.3 Skill scores

The primary skill scores used are Probability of Detection, POD, Confidence, C (with False Alarm Ratio, FAR=1−C), and Critical Success Index (CSI). These skill scores are extended to accommodate near misses (the forecast misses but is within tolerance) and close false alarms (an event is forecast but is nevertheless within tolerance of the observed non-event). The scores are defined as:

$$\mathrm{POD}=\frac{a+n_m}{a+c}, \quad C\equiv 1-\mathrm{FAR}=\frac{a+c_{fa}}{a+b}, \quad \mathrm{CSI}=\frac{a+n_m+c_{fa}}{a+b+c} \qquad (1)$$

where, for a given lead-time window $L \pm \Delta t$, a is the number of times a threshold is crossed by both the modelled and observed series (a *hit*), b is the number of times only the model indicates a threshold crossing (a *false alarm)* and c is the number of times only the observed series indicates a threshold crossing (a *miss).* The final category count, d, is the number of non-threshold crossings occurring in both observed and modelled series (a *correct rejection*). The counts c_{fa} and n_m denote the number of close-false-alarms and near-misses respectively, where modelled and observed threshold crossings occur within some tolerance (see [2, 3] for details). These two variables can be set to zero for the definitions of POD, Confidence and CSI in (1) where a tolerance is not applied.

The POD score gives, if an observed event crosses a threshold, the probability that the forecast will also cross it within some timing tolerance of the observed crossing. The confidence C gives, if a forecast crosses a threshold within a lead-time window, the probability of the observation also crossing it within some timing tolerance. The Critical Success Index CSI is a composite performance measure combining the opposing POD and FAR scores into an overall statistic, with values ranging from 0 to 1 as performance improves.

The performance statistics have been specifically chosen to suit usage in an operational setting. For example, *Confidence* gives the probability of an event (crossing of a threshold in an upward direction) that has been forecast actually occurring, giving useful guidance to the operational user of the forecast. The skill scores can be applied to a specific lead-time, over a full forecast horizon, or over the early, middle and late part of a forecast; and tolerances can be invoked or not. If the Confidence is calculated from a model using forecast rainfall as input, then it might be thought to more accurately reflect operational conditions. However, a model analysed with perfect rainfall has a Confidence that reflects the model performance and not the vagaries of the rainfall forecast and may prove more useful in practice.

Whether the forecast threshold crossing is likely to be late or early is also useful knowledge to the operational forecaster, as is the most likely time of the crossing. The *mean time difference* gives, for all observed events, the average time difference between each forecast and observed event. A histogram of the time difference is constructed giving additional information on whether a particular model tends to forecast events too early or late, an indication of the timing uncertainty, and the time range within which some percentage of events occur.

3.4 Skill score calculation from model forecasts

To make optimal use of the model forecast datasets an ‘all-available forecast’, and not a ‘one-event-per forecast’, approach was followed. Information in every available forecast is used in the construction of the skill scores, rather than just those forecasts that match with an observed event. This approach is aligned to how forecasts are used in practice since, operationally, every forecast is examined and not just those at a specified lead-time before an event (not known at the time to happen). Including all forecasts in a lead-time window provides a more robust measure of model performance at a specific lead-time. Ideally for this approach all forecasts at all lead-times are available for analysis, as is the case for G2G. However, for most local models, forecasts have only been generated prior to known events: the consequence of this is that the False Alarm Ratio will be underestimated for these models.

3.5 Overall Performance Measure

An important objective of the study was to summarise the overall performance of a site model via a single value that could be used in a comparative way at a site or between sites. Although still under review, an Overall Performance Measure (OPM) has been constructed by combining (as an average) the CSI values for the two thresholds and two lead-times of four and 24 hours, the mean timing difference between observed and forecast crossings with and without timing tolerance (scaled to the range 0 to 1), simulation R^2 Efficiency, and the proportion of simulation peaks within magnitude tolerance (±20% peak flow or ±0.2 m peak level) and ±12 hour timing tolerance. The last three measures form the simulation component of the OPM and the first six the forecast component. The latter component, because of its greater operational relevance, is used as the Overall Performance Score (OPS) to rank site models and highlight poorer performing ones. Clearly numerous combinations of behaviour across the individual measures can result in a similar overall performance score. But the OPS serves as a useful quick reference to the general performance of a particular site model for flood forecasting.

4 Analysis of performance statistics

4.1 National evidence-base

Analysis of the model performance statistics was done from both a regional and national perspective to create a national evidence-base useful in strategic planning of forecasting model improvements. The behaviour of model simulations and the success of flood forecasts in different areas of England and Wales were examined in a structured way using a variety of displays. A full report of the analysis with national coverage is given in [3]. Some salient observations drawn from the analysis are summarised below.

- Performance at a lead-time of one hour is the highest of all lead-times for all models.
- Several models, including the PRTF and MCRM, show a marked decline in performance with lead-time. For the MCRM, a mix of perfect and forecast rainfall was used, making it unclear as to how much this is due to the model or to the rainfall input.
- Some models, including the G2G, PDM, TCM and MIKE 11 show a more even performance as a function of lead-time, especially for skill scores including magnitude tolerance. This is likely to be in part an artefact of the larger time tolerances used at longer lead-times. Rather than the model performing better at longer lead-times, it is a reflection of lower expectations of a 'good' forecast at such times.
- Several models - notably MCRM, TCM, MIKE 11 and DODO - have a tendency to forecast events early. For MCRM and DODO this can be attributed in part to the hourly time-step employed, although it is unclear whether this is the dominant factor. PRTF models have good timing of events and high Confidence statistics, but only at short lead times – these forecasts become very poor for lead-times greater than 12 hours.

4.2 The Performance Summary

In addition to the national evidence-base, a primary output of the analysis is a one-page Performance Summary for each site model. This aims to communicate model performance information for both real-time tactical decision-making and offline strategic investment in flood model improvements. The summary contains performance statistics and associated displays, hydrographs for forecast, simulated (if available) and observed levels or flows, and supporting catchment information (e.g. hydrometric network, land cover, artificial influences).

The statistics displayed focus on the success of the model in forecasting threshold crossings via the skill scores described in Section 3. The POD, Confidence and CSI scores, along with the mean time difference, for seven different lead-times are tabulated and displayed in graded form for the two threshold values QMED and QMED/2. The ten biggest observed events over the time period of record analysed, and their associated model forecasts and simulations, are presented as hydrographs. An example from the performance summary is shown in Figure 2. All forecasts within the 48 hour period are shown and colour-coded according to their success at forecasting the peak within a timing/magnitude tolerance box.

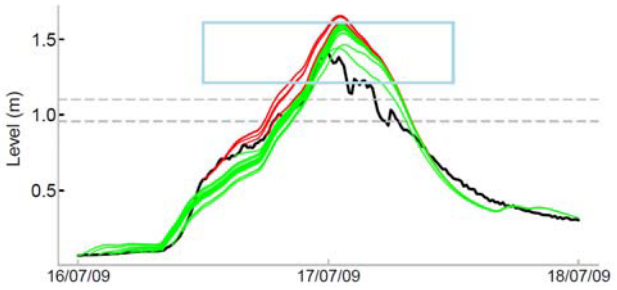

Figure 2. A flood hydrograph from the performance summary for the PDM model at Eastgate on the Rookhope Burn. The black solid line indicates observed flow. Forecasts made every hour are shown and colour-coded green (hit) or red (miss) to indicate their success at forecasting the peak within the blue tolerance box (±12 hour timing and ±20% peak magnitude). Horizontal dashed lines mark QMED and QMED/2.

Figure 3 presents the displays used to jointly summarise timing of a threshold crossing and POD/Confidence skill scores. These displays aim to serve as a direct aid for the interpretation of a flood forecast from a model when used in an operational setting. The time difference histogram, for forecast lead-times of 0 to 36 hours, can be used to assess whether model forecasts are typically early or late, and whether this timing is consistent or variable.

For example, it can be seen from the timing difference histogram that this model (PDM for the Rookhope Burn at Eastgate, a tributary of the River Wear in North East England) forecasts events on time, on average. There is only a slight skew towards earlier forecasts than late and with 90% lying in the range 3 hours early to 2 hours late.

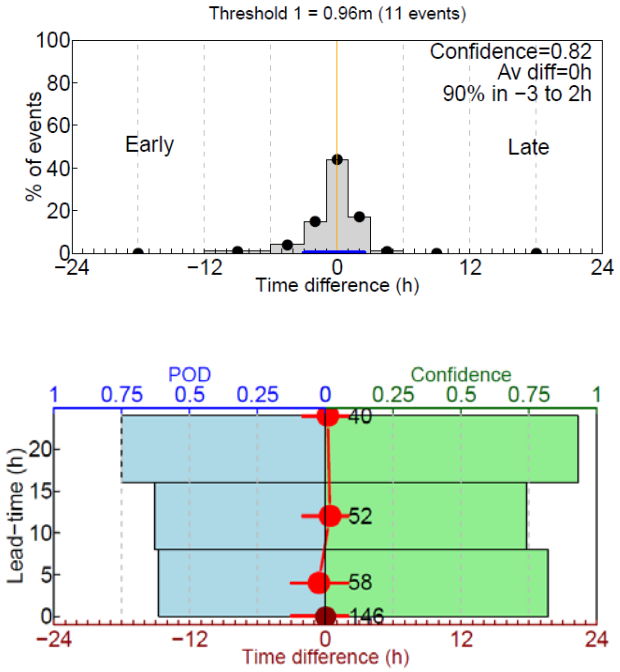

Figure 3. Time difference histogram (top) and forecast performance display (bottom) from the performance summary for Rookhope Burn at Eastgate. A QMED/2 threshold is used.

The lower display presents information on the forecast skill grouped by lead-times of 0-8, 8-16 and 16-32 hours, with the POD score on the left and confidence on the right. In this example, for forecasts with lead-times between 16 and 32 hours there is a 75% chance that an observed event is forecast (POD) and ~90% chance that an event that is forecast by the model actually occurs (Confidence). The red circle and line are positioned to show the time difference as an average and range within which 90% of the forecasts lie. The values indicate the number of events (forecast crossings) within each lead-time range with that for the dark red circle being for all lead-times out to 36 hours.

An important purpose of the performance summary is to allow comparison of performance of two or more site models at the same site. Figure 4 shows an example of the display provided for this purpose. The CSI for a lead-time of 12 hours at the QMED/2 threshold is shown by an open circle for each model at a site, on the vertical axis, whilst the horizontal axis gives the average timing difference.

The closed circle and cross marking the extremities of the vertical line indicate POD and Confidence (equal to 1−FAR) values for the model, whilst the extent of the horizontal line depicts the 90% range of timing differences. The example in Figure 4 indicates that the G2G national model tends to forecast events late whilst the PDM [11] for this same site has good timing, and has higher CSI as might be expected from a locally calibrated model.

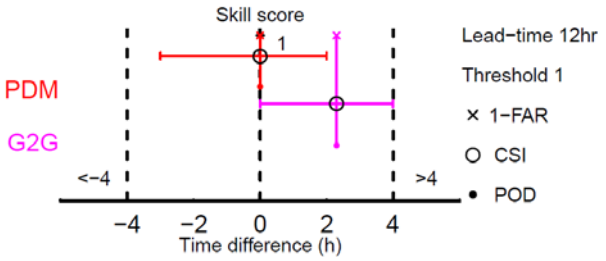

Figure 4. Comparison of model performance display showing skill scores and timing difference of the PDM local model and G2G national model for the Rookhope Burn at Eastgate.

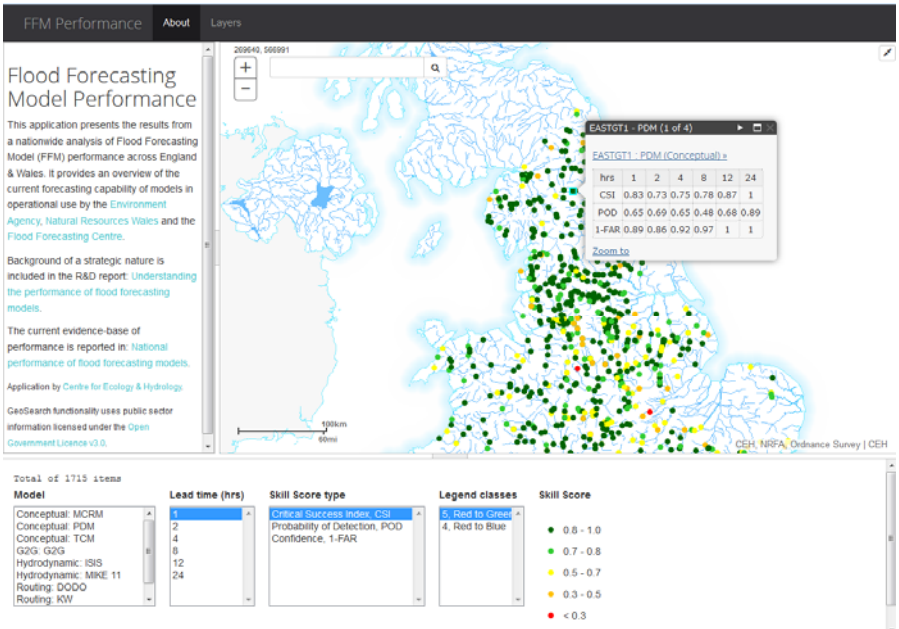

Figure 5. Screenshot of the Flood Forecasting Model Performance web portal. Sites (circles on map) can be filtered by model-type and the performance of one or more model types compared across lead-times for a selected skill score.

4.3 Web Portal

The desire for a more interactive way of interrogating site performance prompted the development of a Flood Forecasting Model Performance web portal, interfacing to the performance summary database developed under the study. A screenshot of the portal is shown in Figure 5. It displays a performance map for sites analysed across England and Wales, which can be filtered by model type, lead-time and performance measure. A pop-up display for a selected site (here for Rookhope Burn at Eastgate) gives a summary of skill scores for different lead-times and the ability to zoom into the location to see the river network and site models in the neighbourhood. The performance summary one page PDF can be accessed from the pop-up display for each site model that exists.

The portal allows new or revised site models to be readily included once they have been processed, easing the rolling out of updates. It is planned to further develop the portal to incorporate more dynamic functionality, easing access to the extensive information contained in the evidence-base and performance summary.

5 Conclusions

A nationally consistent approach for assessing the performance of flood forecasting models has been developed that encompasses both local site-specific models and the national G2G area-wide model. The approach has been applied to a variety of study datasets, collated from past local model performance studies, and to the G2G national dataset, created by running the G2G model offline over five water years to produce forecasts made regularly every hour.

A range of statistics have been chosen to summarise performance at different lead-times and used to analyse forecasting capability by model-type, lead-time and geographic region. The study, reported in detail in [2, 3], has produced an extensive national evidence-base of model performance across England & Wales, and provides an appreciation of comparative performance where models of different type exist for the same site. In addition, a performance summary for each site model has been produced as a PDF using a common one-page template design. It contains performance statistics in numeric and display form that are complemented by hydrographs of forecast, simulated and observed river levels/flows. Some are designed to provide an indication of the success in forecasting an 'event', defined as the upward crossing of a level/flow threshold value. The upper threshold is currently set to have a return period of two years. This can be increased to what may be considered more operationally relevant as the G2G dataset lengthens. Selected displays focus more on the success in forecasting the hydrograph peak. Overall performance measures combine selected statistics to grade each site model. Subsequent ranking of the worst performing site models can be used to troubleshoot and guide future investment in flood modelling and the supporting hydrometric network.

A prototype Flood Forecasting Model Performance web portal has been developed to interrogate the study evidence-base through an interactive map display of forecast performance as it varies with mode-type, lead-time and skill score. Each mapped location also gives access to the performance summary PDF for all site models available at it.

The framework for creating this structured analysis of forecasting model performance has been designed so that site models can be readily updated and added to as they become available. Future updates are envisaged in support of incident management and guiding strategic investment in flood forecasting models across England & Wales. The framework has generic application for assessing and summarising the performance of flood forecasting models across the world.

6 References

1. Werner, M., Cranston, M., Harrison, T., Whitfield, D. and Schellekens, J. (2009). Recent developments in operational flood forecasting in England, Wales and Scotland. *Meteorological Applications*, **16**, 13–22.
2. Robson, A. J., Moore, R. J., Wells, S. C., Rudd, A., Mattingley, P. S and Cole, S. J. (2016). Understanding the performance of flood forecasting models. *Science Report SC130006/R2.* Research contractor: Centre for Ecology & Hydrology, Environment Agency, Bristol, UK.
3. Wells, S. C., Moore, R. J. and Cole, S. J. (2016). National performance of flood forecasting models. *Science Report SC130006/R3*. Research Contractor: Centre for Ecology & Hydrology. Environment Agency, Bristol, UK.
4. Moore, R.J., Cole, S.J., Bell, V.A. and Jones, D.A. (2006). Issues in flood forecasting: ungauged basins, extreme floods and uncertainty. In: I. Tchiguirinskaia, K. N. N. Thein & P. Hubert (eds.), Frontiers in Flood Research, 8th Kovacs Colloquium, UNESCO, Paris, June/July 2006, *IAHS Publ. 305*, 103-122.
5. Bell, V. A., Kay, A. L., Jones, R. G., Moore, R. J. and Reynard, N. S. (2009). Use of soil data in a grid-based hydrological model to estimate spatial variation in changing flood risk across the UK. *Journal of Hydrology*, **377(3-4)**, 335-350.
6. Price, D., Hudson, K., Boyce, G., Schellekens, J., Moore, R. J., Clark, P., Harrison, T., Connolly, E. and Pilling, C. (2012). Operational use of a grid-based model for flood forecasting. *Water Management*, **165(2)**, 65-77.
7. Moore, R. J and Bell, V. A. (2001). Comparison of rainfall-runoff models for flood forecasting. Part 1: Literature review of models. *R&D Technical Report W241*. Research contractor: Institute of Hydrology Environment Agency, Bristol, UK. http://nora.nerc.ac.uk/7471/

8. 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.
9. Wilks, D. S. (2006) Statistical methods in the atmospheric sciences. Second edition, Academic Press.
10. Joliffe, I. T. and Stephenson, D. B. (eds.) (2012) Forecast verification: a practitioner's guide in atmospheric science. Second edition, Wiley-Blackwell.
11. Moore, R. J. (2007). The PDM rainfall-runoff model. *Hydrology and Earth System Sciences*, 11(1), 483-499.