Quantifying and Visualising
Model Performance

About this talk

Some background

Why model performance data is so useful

Measures

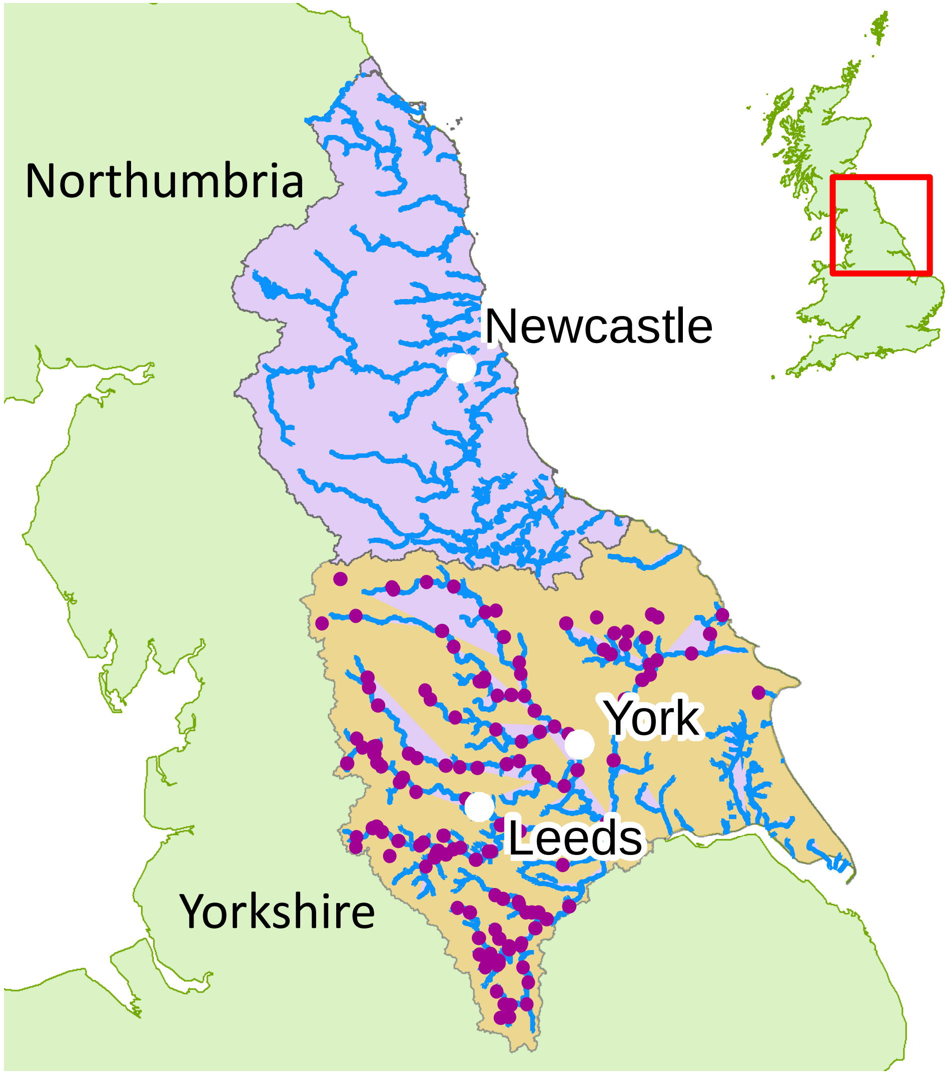
Process

The future

Some background

Quantifying and visualising performance

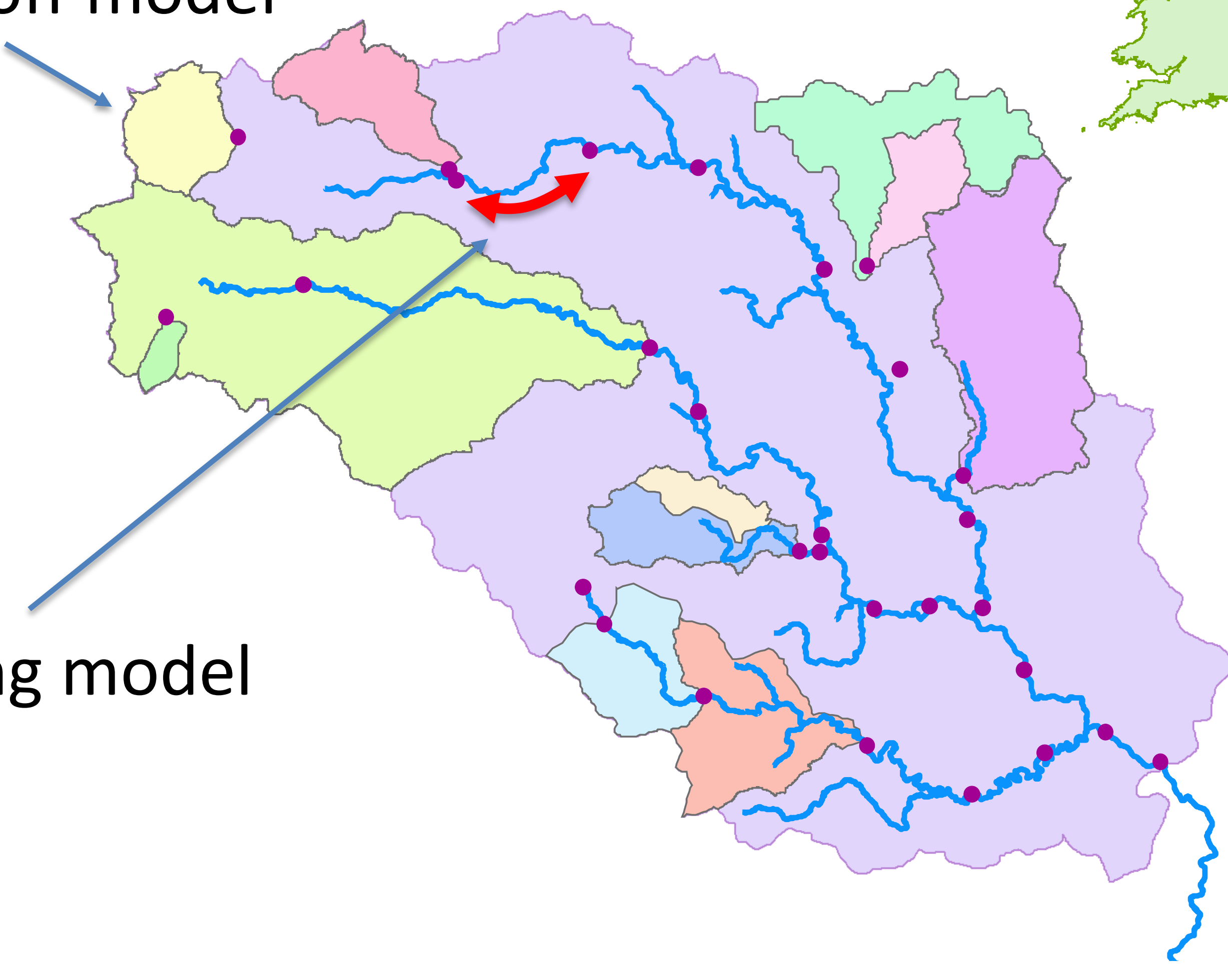
Northumbria and Yorkshire



Ouse: example network

Rainfall runoff model

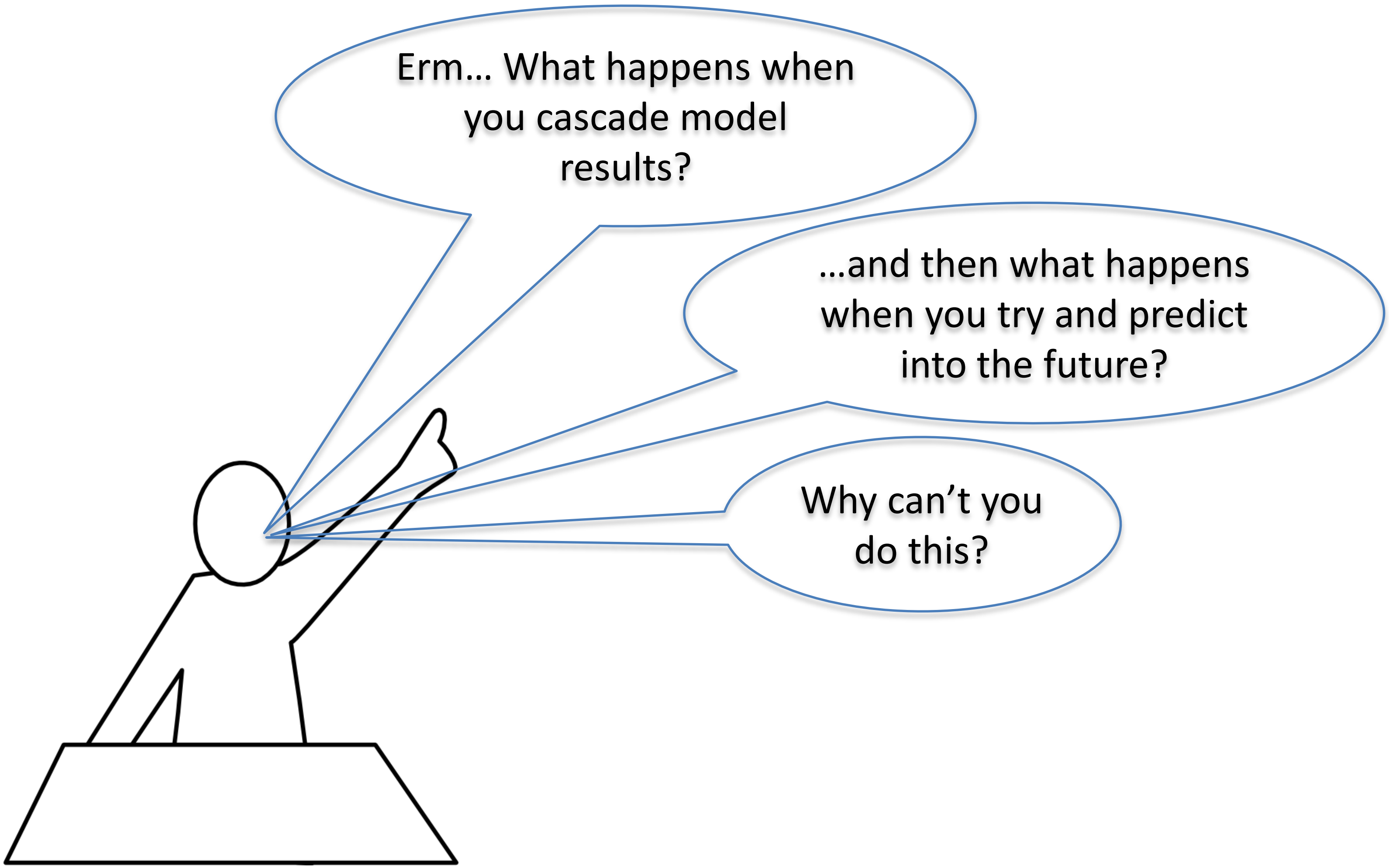
Flow routing model



Model development for RFFS

- Develop individual models
- Calibrate error model
- Supplier integrates
- Hope for the best

Awkward questions



Early attempt

Figure 6.6 : No error correction but
foreknowledge of rainfall

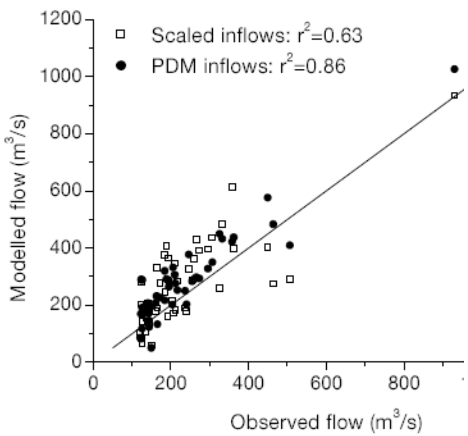
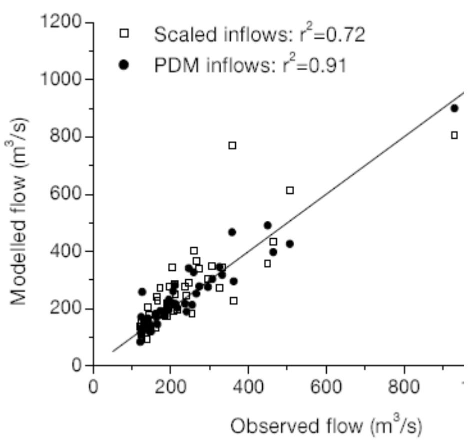
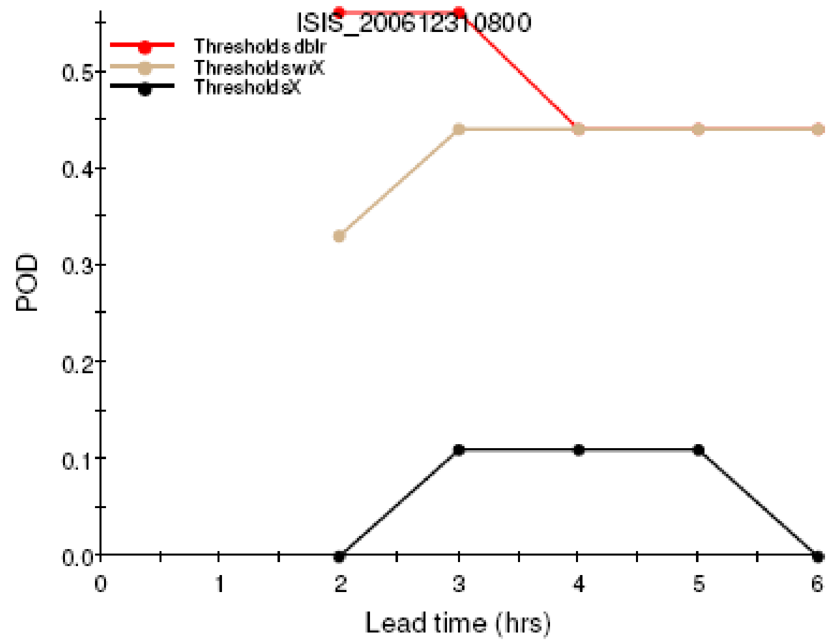


Figure 6.8 : Data available till 3 hours b
peak



Real time performance measures for Colliers



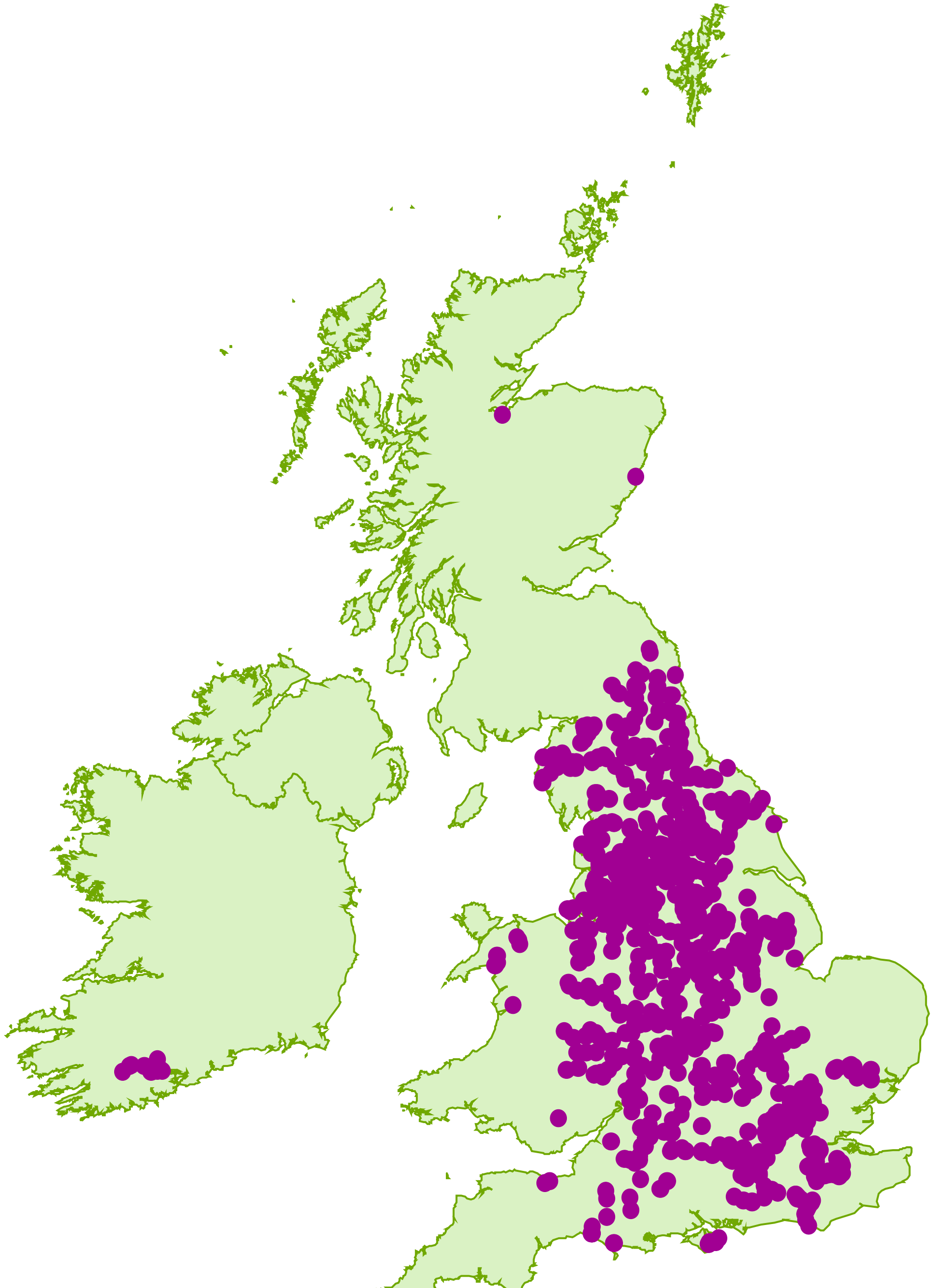
POD table

Scenario	N	POD at increasing lead time (hrs)				
		t-2	t-3	t-4	t-5	t-6
Thresholds dblr	9	0.56	0.56	0.44	0.44	0.44
Thresholds wrX	9	0.33	0.44	0.44	0.44	0.44
ThresholdsX	9	0.00	0.11	0.11	0.11	0.00

And by 2016…

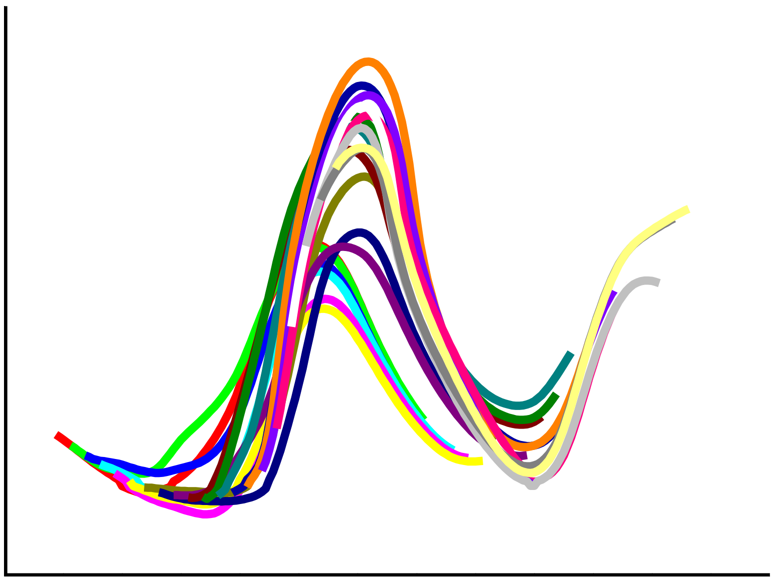
>950 locations tested

Well developed software, skills and methods



Four levels of testing

- Individual component
- Cascaded components
- With data assimilation
- With forecast rainfall

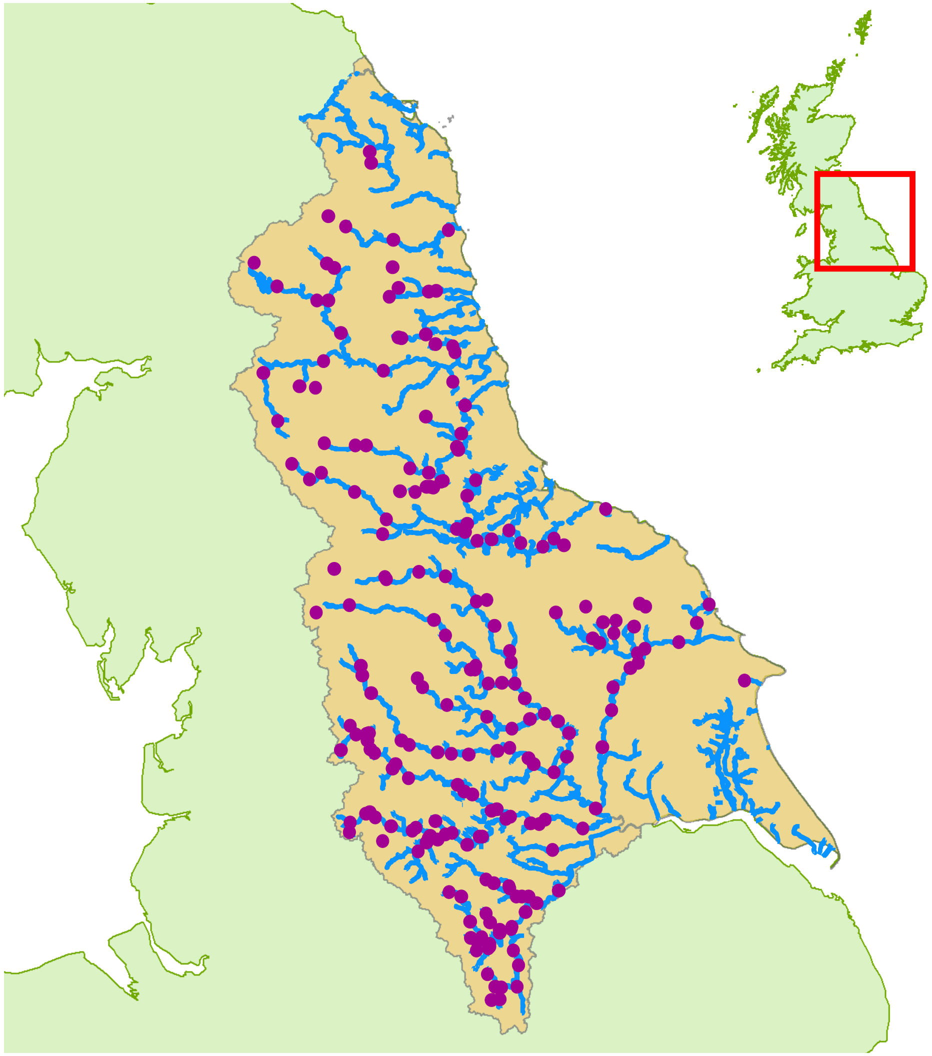


Why performance data are useful

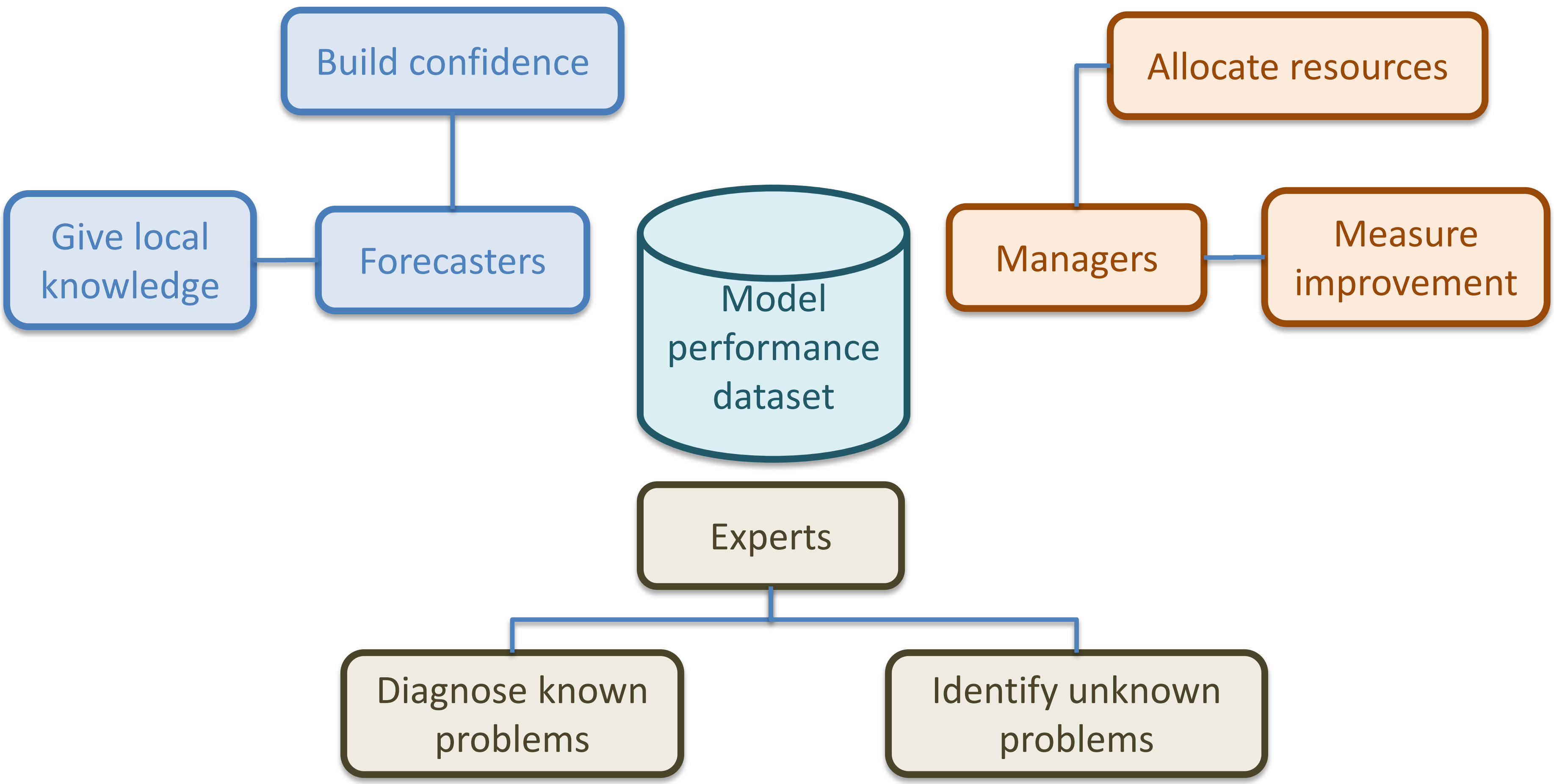
Quantifying and visualising performance

Leeds forecast centre

- Operating since early 1990s
- 210 forecast locations
- 274 models



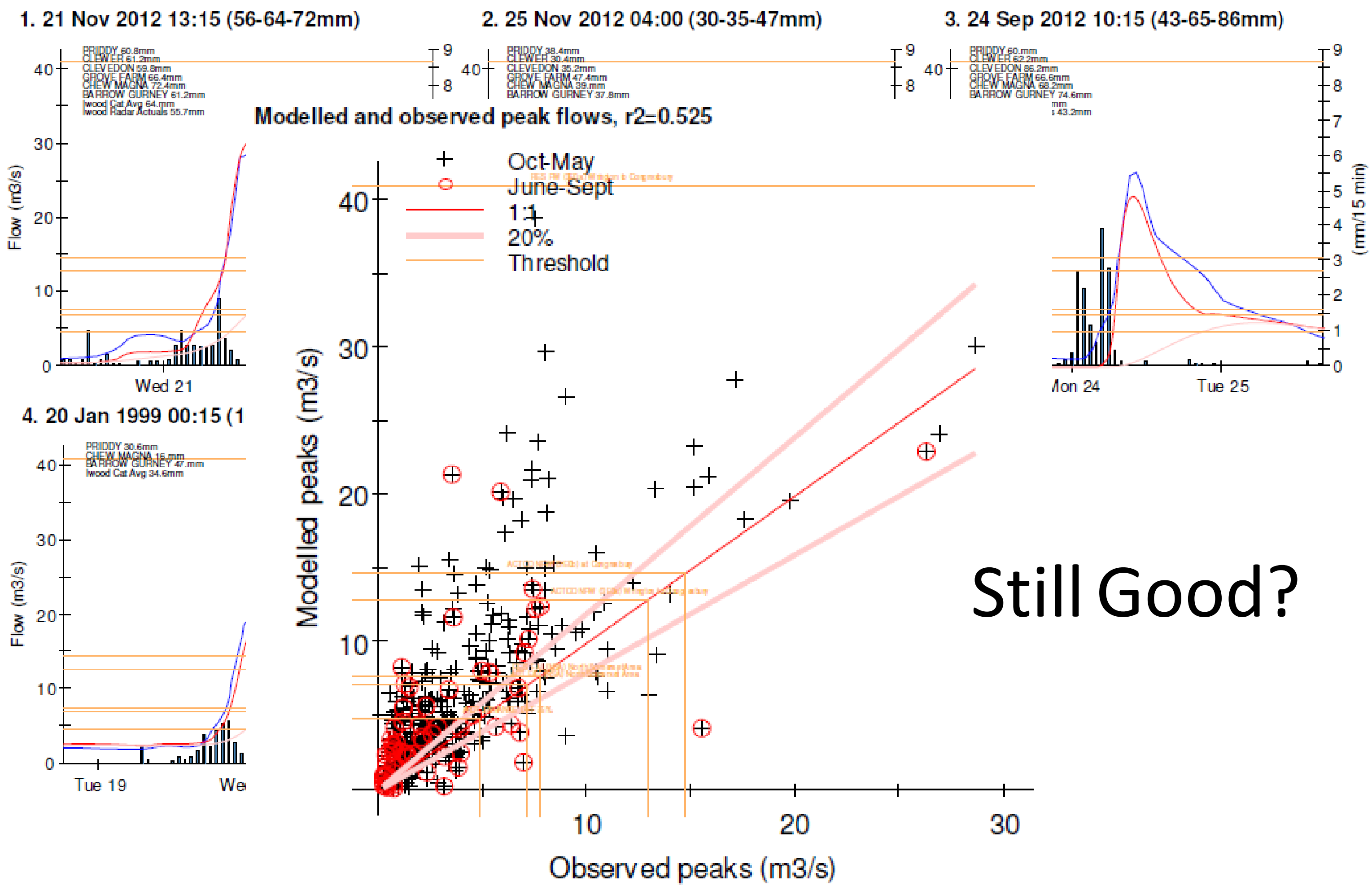
Uses for performance data



How do you quantify performance?

Quantifying and visualising performance

How do you quantify performance?



Still Good?

Types of simulation

- Real time
- Simulation

Questions forecasters are asked

Should we issue a warning – YES OR NO?

Yes

Oh. How big will it get?

Very big

When will it go away?

Soon…

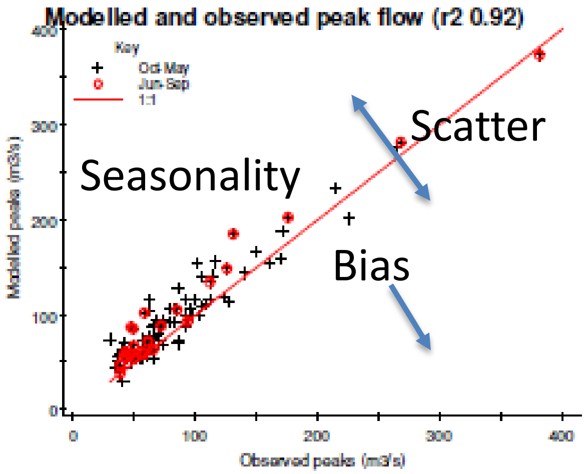
Real time measures

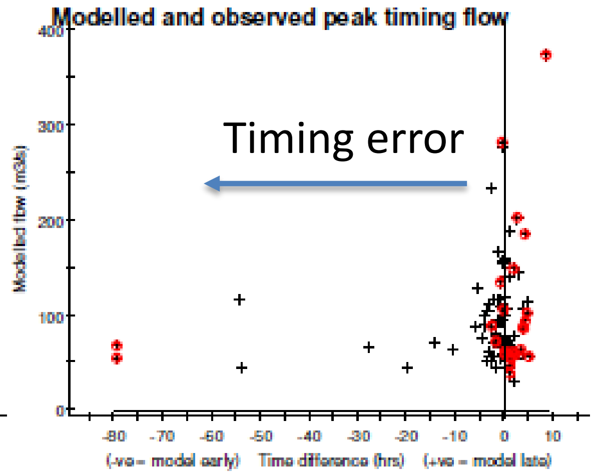
- Probability of Detection (POD)
- False Alarm Rate
- Peak error

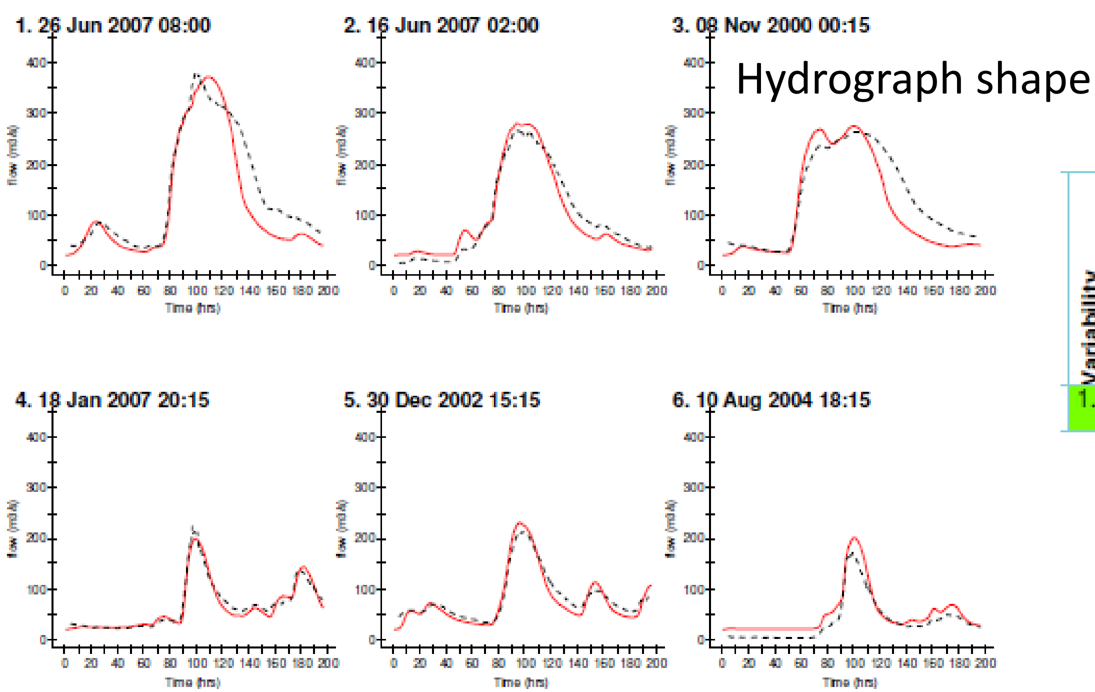
Categorising performance

Grade[2]	Description[1]	Probability of Detection (POD)	False Alarm Rate (FAR)
A+	Exceeds targets	POD ≥ 0.8	FAR ≤ 0.2
A	Meets targets	0.8 > POD ≥ 0.7	0.2 < FAR ≤ 0.3
B	Meets targets with tolerance	POD ≥ target with ±0.2m tolerance	FAR ≤ target with ±0.2m tolerance
C	Does not meet target	0.7 > POD ≥ 0.5	0.3 < FAR ≤ 0.5
D	Significantly below target	0.5 > POD ≥ 0.3	0.5 < FAR ≤ 0.7
E[2]	Poor	POD < 0.3	FAR > 0.7

Simulation measures

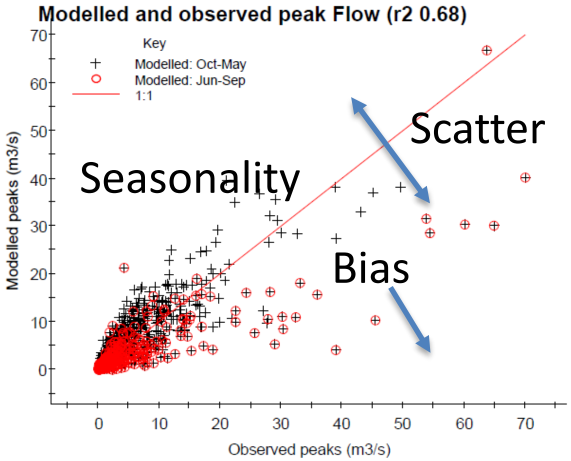


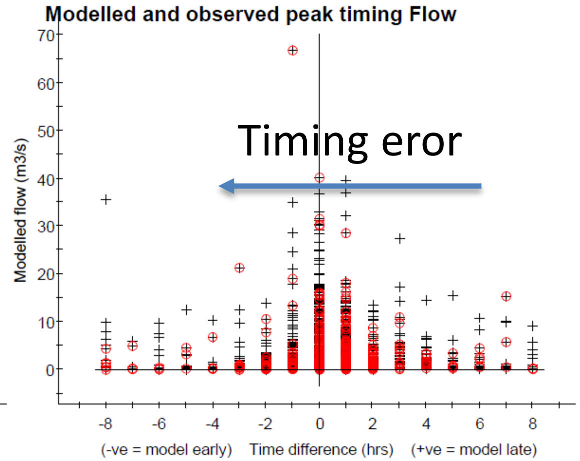


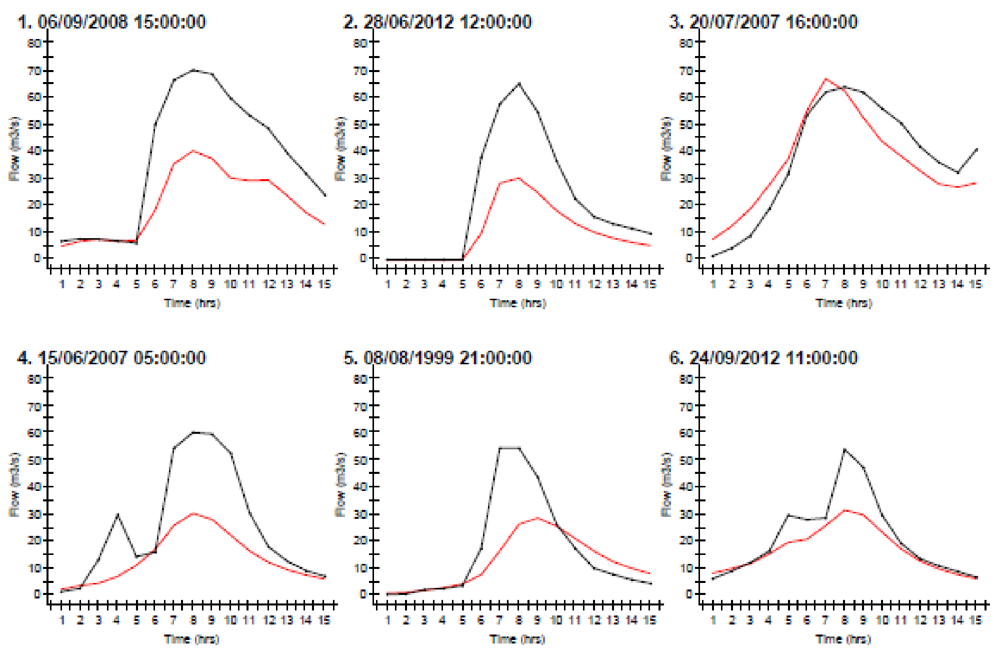


Variability	bias -absolute	bias -significance	shape	Timing bias - abs.	Timing bias - sign.	Seasonaility	Overall Score
1.19	2.62	0.71	1.56	0.58	0.86	2.53	1.36

Simulation measures







Variability	bias -absolute	bias -significance	shape	Timing bias - abs.	Timing bias - sign.	Seasonaility	Overall Score
2.79	3.49	3.54	1.38	2.61	4.00	4.00	2.97

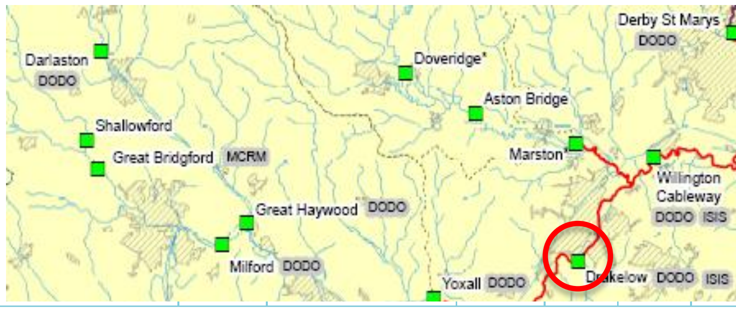
Visualisation

Forecast location performance summary

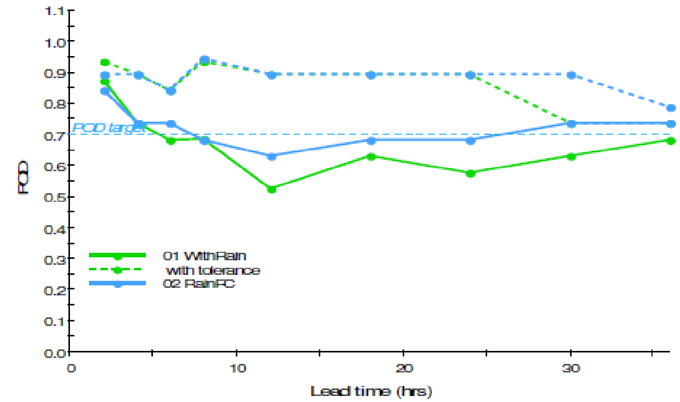
Location ID:	4019
Network:	Middle Trent
Model Type:	ISIS

Real time forecast performance gra

Scenario	Thresh	N
Jan 04 - Jul 14	(m)	
01 WithRain	2.66	
	2.7	
	2.88	
	3.25	
	3.45	
	3.8	
	Peaks	
Jan 04 - Jul 14	(m)	
02 RainFC	2.66	
	2.7	
	2.88	
	3.25	
	3.45	



Real time performance measures for 40192.66H.rated.forecast



Model scores

	Variability	bias -absolute	bias -significance	shape	Timing bias - abs.	Timing bias - sign.	Seasonality	Overall Score
Modelled	3.26	3.37	3.17	1.46	2.15	3.17	4.00	2.93

D table; threshold tested = 2.66m

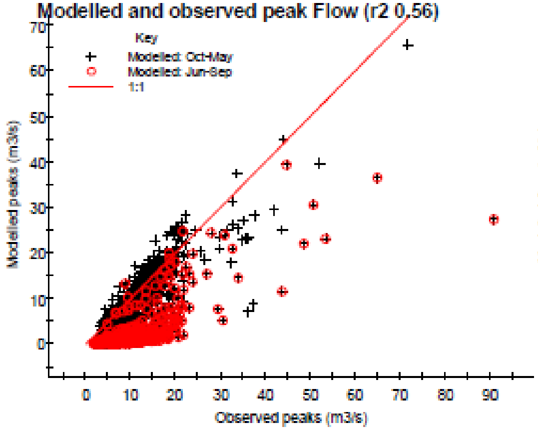
enario	N	POD at increasing lead time (hrs)								
		t-2	t-4	t-6	t-8	t-12	t-18	t-24	t-30	t-36
WithRain	19	0.88	0.74	0.68	0.69	0.53	0.63	0.58	0.63	0.68
RainFC	19	0.84	0.74	0.74	0.68	0.63	0.68	0.68	0.74	0.74

the number of observed crossings used to calculate POD
olute POD values are tabulated. POD values with a tolerance of 0.2 are also plotted as dashed lines, i.e. if a
ulation gets within a specified distance of the threshold, then it counts as a hit

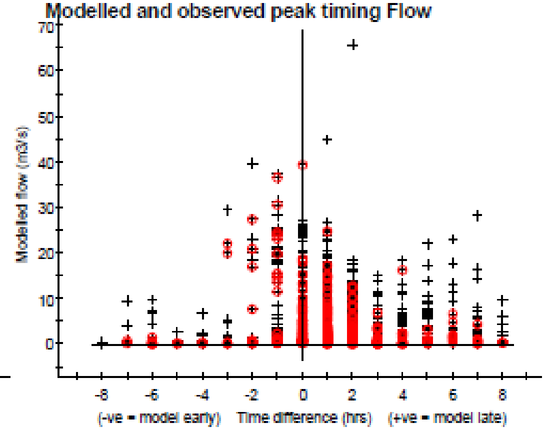
WITH upstream error correction

1.1	0.6	0.2	1.5	2.4	3.8	1.7	1.5
1.3	2.5	0.5	0.6	2.3	2.2	2.8	1.5
2.1	4.0	2.1	1.0	1.5	1.1	4.0	2.1
1.5	0.2	0.1	1.2	0.1	0.1	2.2	1.0
1.6	3.3	2.1	2.2	2.2	2.7	3.6	2.3
1.5	2.4	0.8	1.8	3.4	3.8	4.0	2.3
1.0	4.0	2.3	1.0	0.7	1.0	2.0	1.4
3.3	3.4	3.2	1.5	2.2	3.2	4.0	2.9
3.0	3.0	2.5	1.6	1.9	3.8	1.1	2.4
2.2	0.7	0.4	1.2	1.8	3.9	4.0	2.1
1.9	2.9	2.4	2.7	1.4	3.5	2.9	2.5
1.0	2.7	0.6	0.8	1.8	1.9	2.2	1.3
2.6	2.3	1.8	2.4	1.1	4.0	2.3	2.4
1.6	1.1	0.5	1.8	0.8	0.7	2.3	1.4
1.3	1.4	0.6	2.1	2.4	4.0	1.6	1.8
2.7	0.4	0.7	1.5	0.5	2.3	4.0	2.1
2.5	1.7	2.1	1.9	0.2	0.5	3.5	2.0
2.8	3.5	3.5	1.4	2.6	4.0	4.0	3.0
1.1	4.0	2.4	1.4	1.5	2.9	2.7	2.0

Modelled and observed peak Flow (r2 0.56)



Modelled and observed peak timing Flow



Explanation of performance grades
Explanation of model scores

Visualisation

- Forecasters

Real time forecast performance grades

Scenario	**Thresh.**	**N**	**Lead time (hrs) [grey cell = indicative catchment response time]**									
Jan 04 - Jul 14	**(m)**		**2**	**4**	**6**	**8**	**12**	**18**	**24**	**30**	**36**	
01 WithRain	2.66	19	B	B	A	A	A	A+	B	B	B	
	2.7	16	A+	A	A+	A+	A+	B	B	B	B	
	2.88	10	A+	A+	A	A+	A+	A+	B	B	B	
	3.25	5	B	B	B	B	B	B	B	B	E	
	3.45	2	B	B	B	B	B	B	B	B	C	
	3.8	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Peaks		A	A	A	A	A	B	B	B	A	
Jan 04 - Jul 14	(m)											
02 RainFC	2.66	19	B	B	A	A	A+	A+	A	B	B	
	2.7	16	A+	A	A+	A+	A+	A+	B	B	B	
	2.88	10	A+	A+	A	A+	A+	A+	A	B	B	
	3.25	5	B	B	B	B	B	B	B	E	E	
	3.45	2	B	B	B	B	B	B	B	B	C	
	3.8	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
	Peaks		A	A	A	A	A	B	B	B	A	

Visualisation

- Managers

See footnote for an explanation of table column headings

Forecast location performance commentary

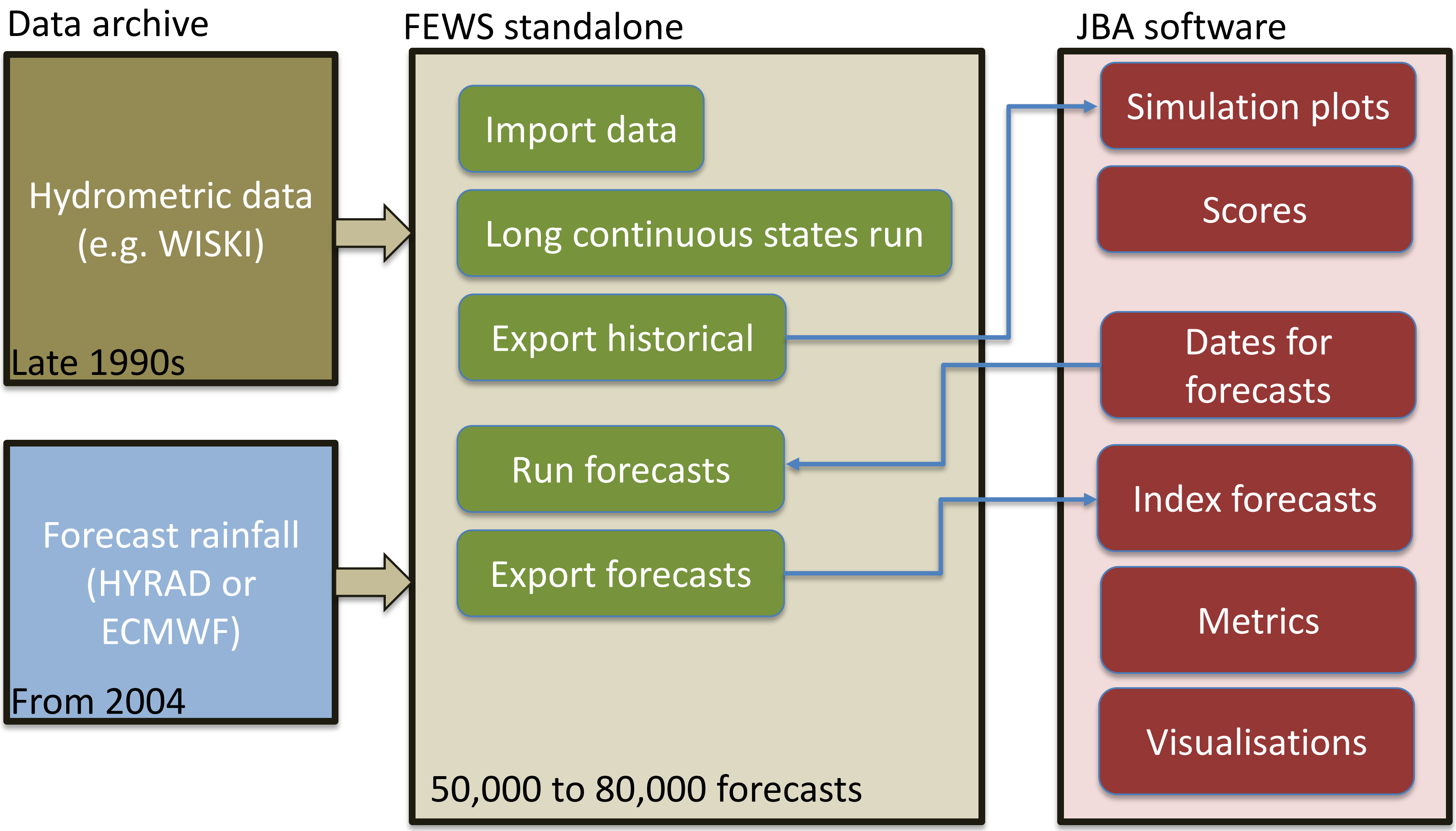
JBA consulting

ID	Site	River	Type	Response time	Obs Rain	Fcast Rain	Intrinsic	Variability	Bias abs	Shape	Timing bias abs	Seasonality	Overall Score	Overview	Recommendations
2516	Halesowen	Stour	MCRM	2	B	B	*D*	2.2	3.1	3.4	0.7	3.6	2.8	MCRM simulating the headwaters of the Upper Stour. Catchment urbanised	Consider modelling the urban response separately to improve the model's performance in smaller events particularly.
2083	Stourbridge	Stour	DODO	2.31	B	B	E	3.0	2.3	1.5	1.2	2.9	2.5	DODO routed flows from Halesowen, with significant urban lateral inputs from an MCRM	For lateral inlfows, consider modelling the urban response separately to improve the model's performance in smaller events particularly. Calibrating the model's wavespeed may improve performance above 20m3/s.
2641	Wightwick	Stour	MCRM	2	B	B	*E*	2.4	1.0	3.4	1.2	0.4	2.2	Small MCRM simulating the Upper Smeston. Fast response	Consider recalibration to improve the hydrograph shape (this will also require a reduction in volume of runoff)
2706	Wombourne	Stour	MCRM	2	A	D	*E*	3.0	2.0	3.7	2.0	4.0	3.2	Small MCRM simulating the Wom Brook. Fast response, with an obviously early urban runoff peak	Consider modelling the urban response separately to improve the model's performance in smaller events particularly.
2067	Swindon	Stour	DODO	4	A+	*B*	C	1.8	1.4	1.4	2.2	0.2	1.9	DODO routes flows from Wombourne and Wightwick with additional lateral inflow from an MCRM. Observed hydrographs quite 'flat topped', indicating storage or bypassing	Understand the cause of the 'flat topping' in the observed series and decide whether it needs to be incorporated into the DODO model
2084	Stourton	Stour	DODO	8.5	A+	B	*B*	1.2	3.2	1.1	1.0	3.6	1.8	DODO routing flows from Swindon and Stourbridge with some lateral inflow from an MCRM	Consider correcting the peak bias and timing bias by adjusting the DODO's inputs and wavespeed (although low priority)
														DODO routing flows from Stourton with additional lateral flow from MCRM	Investigate the source of the bias and then correct.

How can I get my hands on this stuff?

Quantifying and visualising performance

What's the process?



What specialist skills do I need?

- FEWS configuration
- Software for reporting/processing
  - but could use FEWS performance module

The future

Quantifying and visualising performance

In Future?

- Testing now the norm
- National results database
- Changing measures
- Make it easier to re-run forecasts (archive)
- Other uses for results (e.g. QR)

Thanks for listening!

Quantifying and visualising performance