ISBN number for HWRS 2021 is 978-1-925627-53-4

Machine learning to improve hydrologic calibration and reduce risk

Kyle Thomson

Water Modelling Solutions

kyle.thomson@watermodelling.com.au

ABSTRACT

The calibration or validation of a hydrological model is critical in the process of calculating design flows. The consequence of underestimating parameters during the validation process can result in an underrating of the flood risk, and thereby the negative impact on infrastructure costs, insurances, population at risk and other indirect effects.

There is no universally recognised best practise or definitive standard for hydrological modelling because each calibration is unique and carries its own challenges and limitations. This means that the hydrologist is required to exercise significant judgement with regard to the selection of an approach and the parameters for the estimation process.

To mitigate the impact of this subjective bias, WMS has developed a machine learning algorithm with an ensembled approach that groups favourable parameter values and thereby identify and disqualify outliers. This paper describes the results of a pilot study that was conducted on five (5) large catchments in Queensland with the aim of varying the quality of the recorded data to test the efficiency of the machine learning algorithm.

The initial results are positive when they are benchmarked against a performance criterion developed by WMS, which is based on Nash–Sutcliffe efficiency coefficient (Nash, J. E.; Sutcliffe, J. V. 1970), Kling-Gupta efficiency coefficient (Gupta et al. 2009) and others. The results of this study further emphasise the importance of including base flow, spatially varying rainfall, consideration for hyetograph shape and volume-duration in the calibration process.

INTRODUCTION

Hydrological modelling is stochastic in nature. When calibrating a model this behaviour becomes more apparent as the uncertainty and scale of the inputs increases. Inputs during the calibration process that influence this uncertainty generally include the following:

- The catchment delineation;
- Rainfall hyetographs and spatial distribution in the catchment;
- Loss model;
- Baseflow or baseflow separation;
- Accuracy of the stream gauge;
- Catchment characteristics; and
- Modellers bias in the calibration process

The objective of this paper is to develop a methodology that can help to reduce modellers' bias using machine learning (ML). The ML algorithm is itself based on a set of assumptions which may be regarded as identified biases (see *methodology*) which the author of this paper has attempted to reduce by introducing a broad performance criterion.

Similarly to hydrology, market algorithmic trading (algo-trading) is highly stochastic, where predicting the change in behaviour over time is hard to capture with any degree of certainty. A popular ML technique (ensemble-learning) in the algo-trading community was adopted for this study.

The results presented in this paper have been developed with sole dependence on the use of ML and minimal modellers judgement. The author does not endorse the use of any results presented for the gauges used in this paper.

METHODOLOGY

Five catchments were selected in Queensland with varying catchment characteristics, size of areas and quality of data. They were selected with the following criteria:

- Stream and rainfall gauges in close proximity with data available publicly;
- Extensive stream and rainfall overlapping data; and
- Varying total areas across the five catchments

Catchments were generated, delineated into sub-catchments, reaches and nodes using CatchmentSim (CSSE 2021) with SRTM-H data (Geoscience Australia 2006).

The event based hydrological model RORB was chosen given the simplicity of input parameters compared to other models. Assuming a fixed rainfall, unregulated catchment (no water storage structures) and an m value of 0.8, a RORB model can be calibrated varying three constants. These are initial loss (IL), continuing loss (CL) and a fixed routing parameter (Kc).

Where possible, baseflow was separated from stream hydrographs to improve the machine learning performance. Generally this was only done when the ML algorithm couldn't achieve a calibration (refer to *Estimate of Kc and CL values*).

Rainfall was only spatially varied for the largest catchment of the five (136301B), this was to test the ML algorithm performance with a higher quality of data for a large catchment.

Rainfall was not manipulated in any way except for gauge 136301B where it was spatially distributed. No interpretation of rainfall was made to remove pre or post storm bursts, rather the ML algorithm was left to select an IL that would allow for a best fit to various performance criterions. Due to this approach taken, no ILs described in this paper would be suitable for their catchments design hydrology.

A multi-layer perceptron regressor (MLP-R) neural network solver was used with the python module scikit-learn (Figure 1), MLP-R is a method of supervised learning where a function is developed by training on a dataset. The training datasets used were generated by running each hydrological model for a sequence of random inputs and compiling them with the resulting hydrographs.

A suite of performance criterions was incorporated into the MLP-R to compare the modelled (also referred to as testing) hydrographs to the recorded stream gauges. These were:

- Nash–Sutcliffe (NSE) efficiency coefficient (Nash, J. E.; Sutcliffe, J. V. 1970). Used for comparison of hydrograph and volume-duration timeseries data;
- Kling-Gupta (KGE) efficiency coefficient (Gupta et al. 2009). Used for comparison of hydrograph and volume-duration timeseries data;
- Pearson correlation coefficient. Used for comparison of volume-duration timeseries data; and
- Percentage difference in hydrograph peak flow.

An ensemble of the MLP-R network was developed with a randomised range of IL, CL and Kc inputs per network. Testing results for each network were grouped and benchmarked with the performance criterions. Outliers removed where applicable. An example of a testing result to one of the trained simulations for this paper is shown in Figure 2.

This process was repeated (iterated) until a range of inputs that perform well against the performance criterions was achieved. If the performance criterions could not be achieved the ML algorithm eventually failed. A visual workflow of the process is shown in Figure 3.

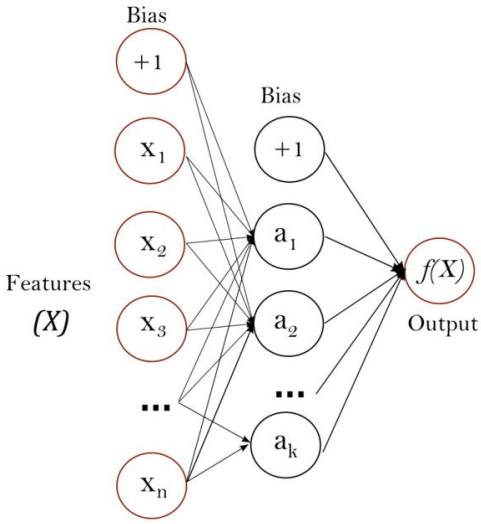

Figure 1. One Hidden Layer MLP (scikit-learn, 2021)

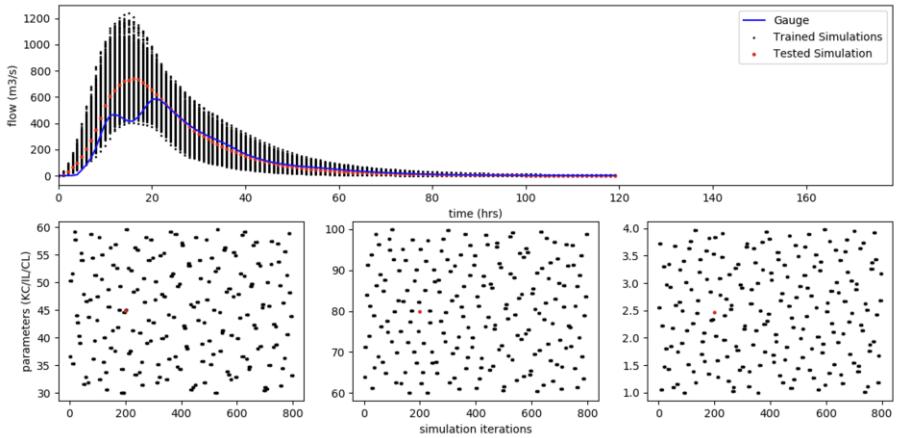

Figure 2. RORB ML Algorithm Tested Result

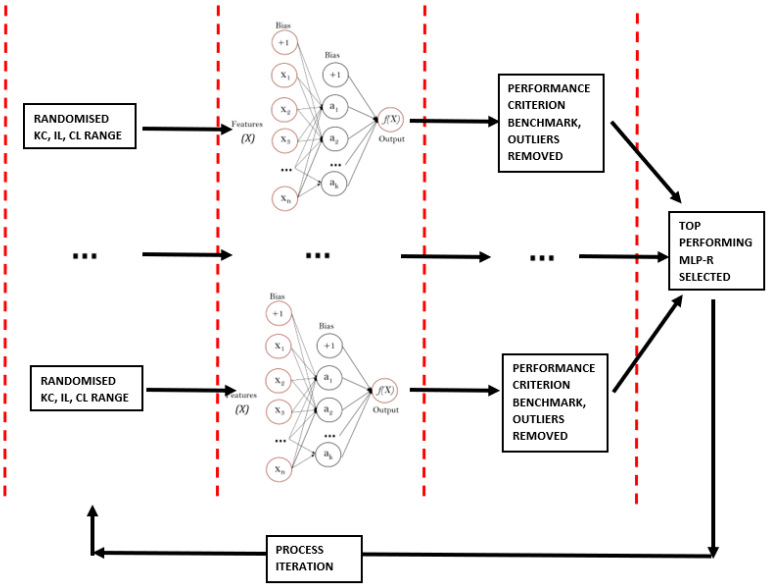

Figure 3. Ensembled-learning Process

PILOT GAUGES

Stream gauges were selected in Queensland as part of this pilot study. The total contributing catchment area of these gauges varied between 35 to 500 km^2. A summary of the selected gauges is shown in Table 1.

Kc values traditionally derived for uncalibrated catchments in Queensland are also shown in Table 1. These values are from the default RORB equation, Weeks (1986) and Aus Wide Dyer (1994; Pearse et al., 2002). The continuing loss (CL) at each gauge was also extracted from Datahub (Babister et al, 2016) as a reference before the ML calibration.

For gauge 125006A, pluviograph 1250P002 was selected due to its proximity to the catchment centroid. Gauges 125009A, 126003A and 141009A had pluviograph data at the stream gauge locations, these were used due to overlapping historical periods. One of these gauges and its contributing catchment is shown in Figure 4.

Gauge 136301B didn't have a pluviograph of sufficient quality within the catchment. Due to the catchment size many hourly and daily pluviographs were used with a spatial distribution applied to better estimate rainfall depths across the catchment. This gauge and its contributing catchment is shown in Figure 5.

Table 1. Pilot gauges and typical parameters

Gauge	Area (km^2)	Average flow distance (km)	Kc (Default)	Kc (QLD, Weeks)	Kc (Aus Wide Dyer)	CL (Datahub)
141009A	43.0	6.55	14.42	6.46	7.47	2.6
125006A	35.4	3.64	13.10	5.83	4.15	5.2
125009A	192.3	10.97	31.51	14.29	12.5	5.2
126003A	83.6	8.33	20.11	9.19	9.49	2.4
136301B	494.5	28.41	48.92	23.57	32.39	1.5

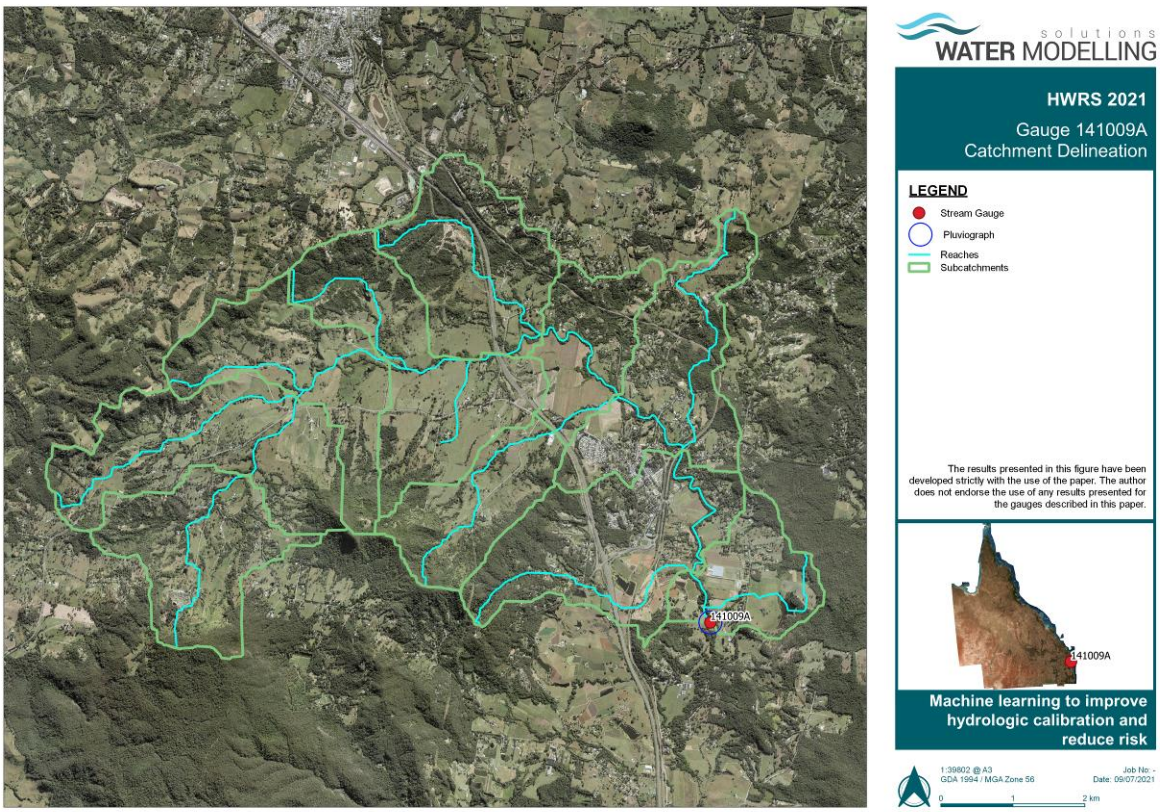

Figure 4. Gauge 141009A Catchment Extents

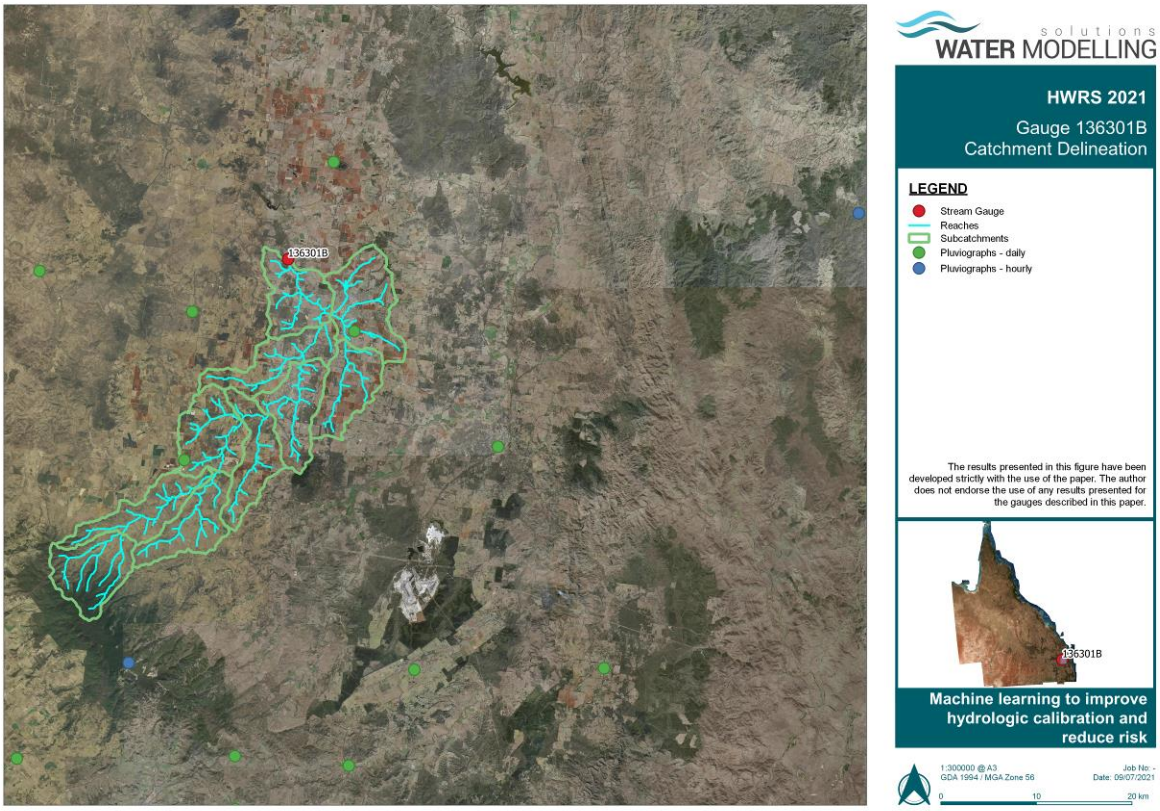

Figure 5. Gauge 136301B Catchment Extents

ESTIMATE OF KC AND CL VALUES

Where data of sufficient quality existed, the five rarest flood events from each gauge were extracted and run through the ensemble-learning process. Results are summarised in Table 2 to Table 6 with commentary on the findings.

Model outputs of the final iteration of each event with plots of the each resulting performance criterion is shown throughout this section. For the performance criterion plots, "model convergence" describes the final iteration where the best fit was achieved for the given event.

Table 2. 141009A - Ensemble-learning model results

Event	Kc Range	CL Range	Commentary
Feb 2012	20.1 – 22.1	1.8 – 2.6	Good fit to hydrograph and volume shape was achieved, difference in time to peak is likely due to the pluviograph not capturing the catchment average hyetograph accurately.
Jan 2013	19.7 – 21.5	4.2 – 4.9	Close to perfect fit when benchmarked to the performance criterion.
Feb 1999	25.2 – 25.2	3.6 – 3.6	Visual observation suggests insufficient volume from the pluviograph impacting results.
Mar 1997	24.3 – 25.8	5.6 – 6.4	Good fit to hydrograph and volume shape was achieved, difference hydrograph shape is likely due to the pluviograph not capturing the catchment average hyetograph shape accurately.
Apr 2009	15.1 – 16.0	2.5 – 3.5	Visual observation suggests insufficient volume at the start and end of the hydrograph, possible reasons are either from insufficient pluviograph volume or baseflow impacting results.

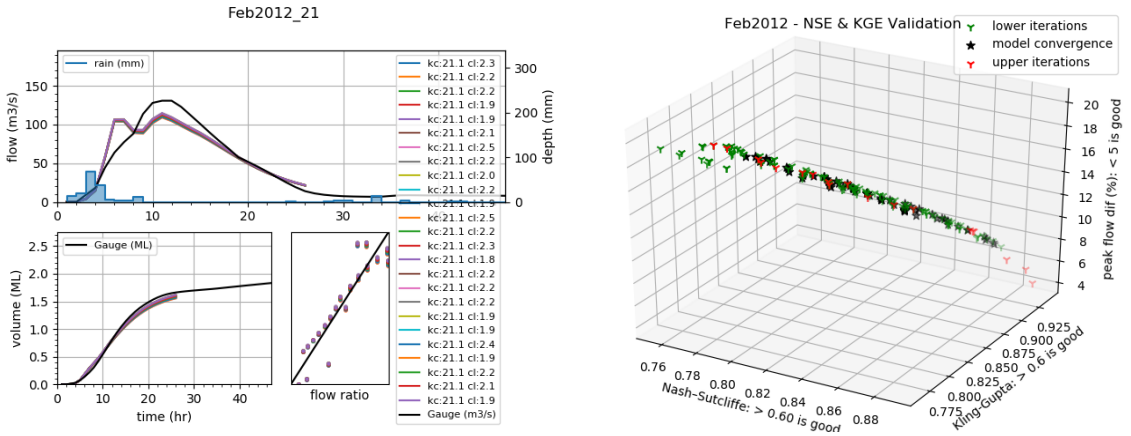

Figure 6. 141009A – Feb 2012 Calibration Results

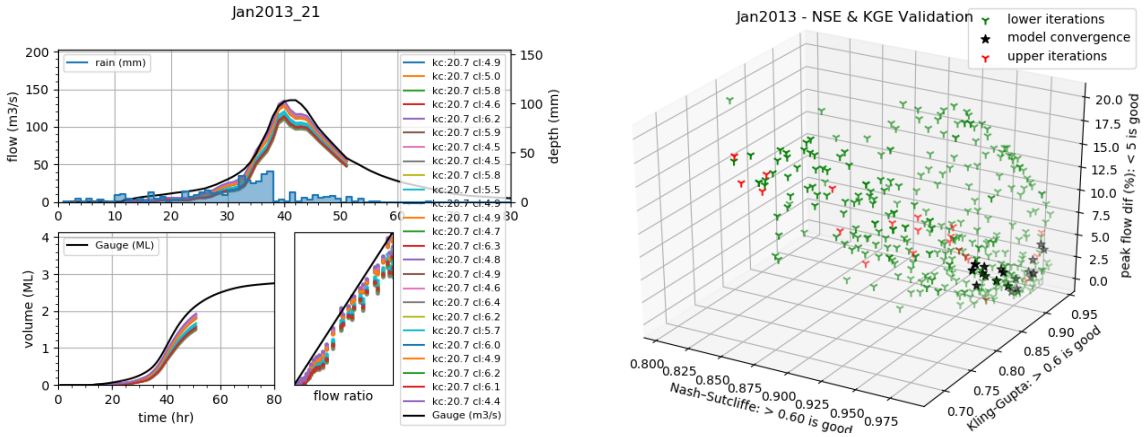

Figure 7. 141009A – Jan 2013 Calibration Results

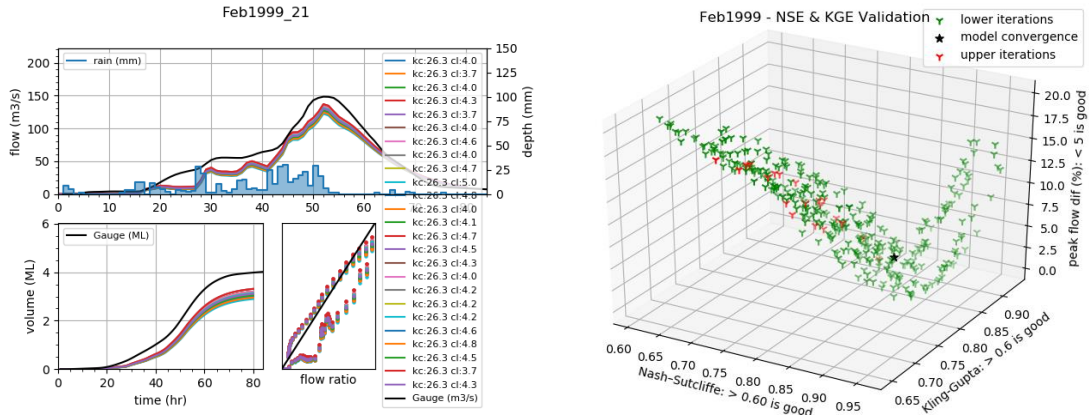

Figure 8. 141009A – Feb 1999 Calibration Results

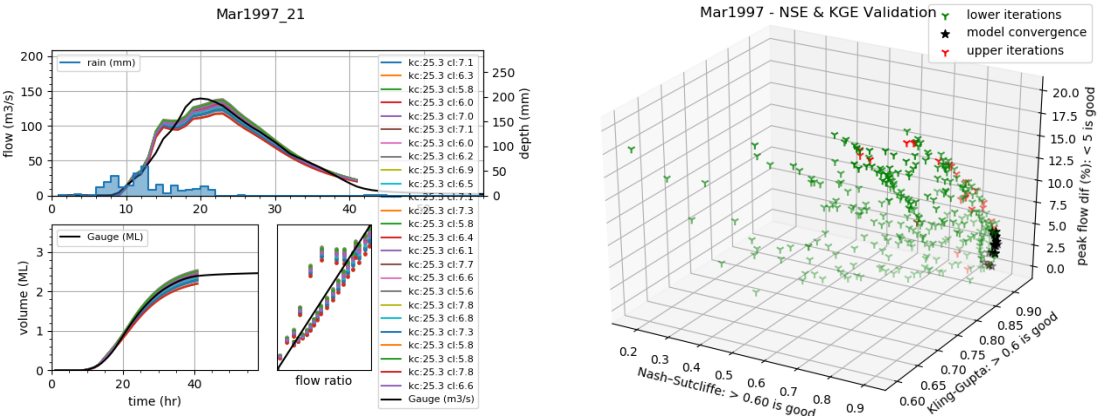

Figure 9. 141009A – Mar 1997 Calibration Results

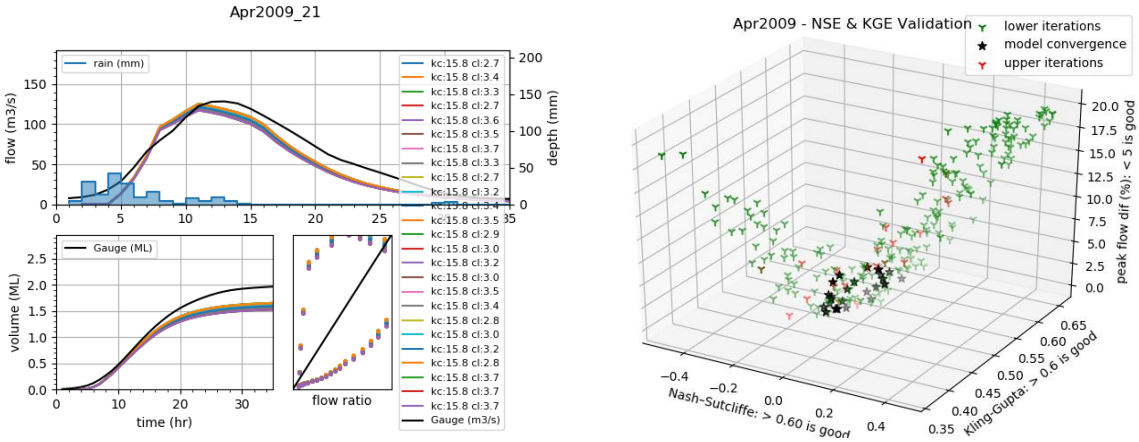

Figure 10. 141009A – Apr 2009 Calibration Results

Table 3. 125006A - Ensemble-learning model results

Event	Kc Range	CL Range	Commentary
Feb 2000	7.0 – 7.6	1.0 – 1.4	Good fit to volume-duration was observed. Hydrograph shape suggests the catchment average hyetograph was not captured accurately.
Jan 2007	7.0 – 7.6	3.8 – 5.3	Notable difference in peak flows. This is likely an over correction from the model due to inaccurate hyetographic information or influence from baseflow.
Aug 1998	6.3 – 6.8	4.0 – 5.4	Poor calibration when compared to the resulting performance criterion. Likely due to storm volume, hyetograph shape and/or baseflow.
Jan 2008	7.3 – 7.8	4.8 – 6.0	A good fit to hydrograph shape and volume-duration was achieved. NSE and KGE suggests a poor fit, likely due to a time lag between the pluviograph and gauge that wasn't captured.
Mar 2017	-	-	ML algorithm failed during calibration

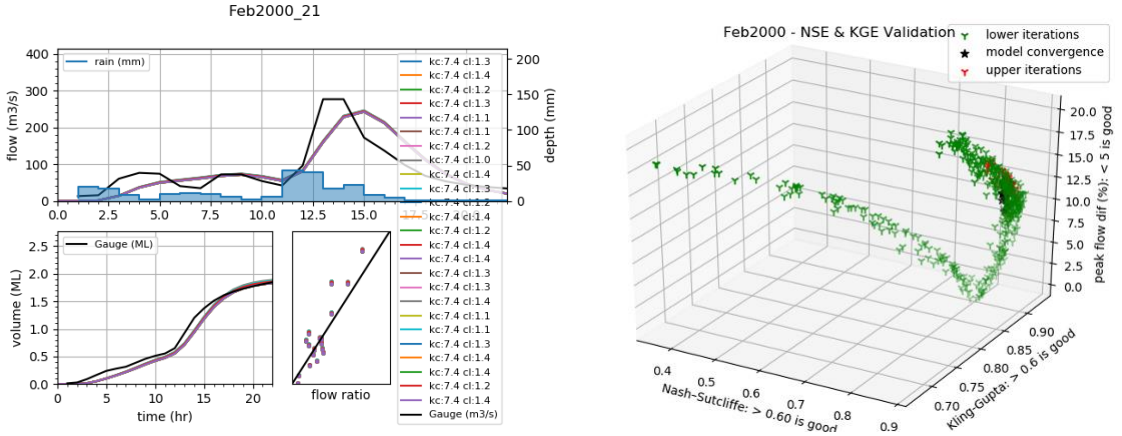

Figure 11. 125006A - Feb 2000 Calibration Results

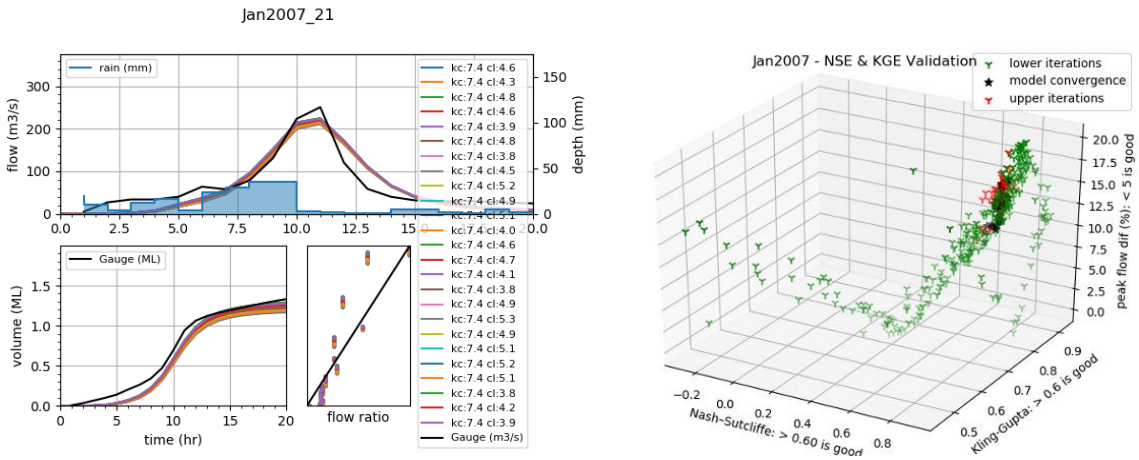

Figure 12. 125006A - Jan 2007 Calibration Results

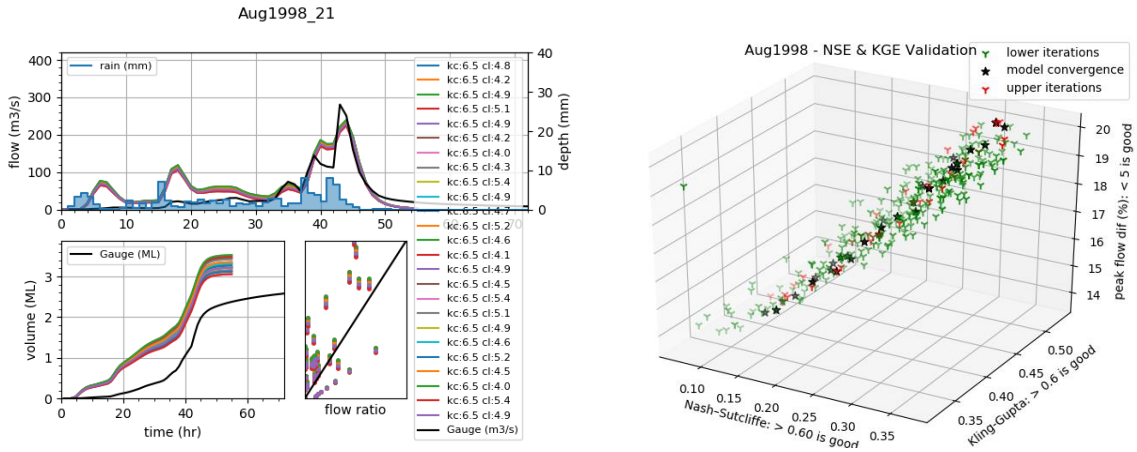

Figure 13. 125006A - Aug 1998 Calibration Results

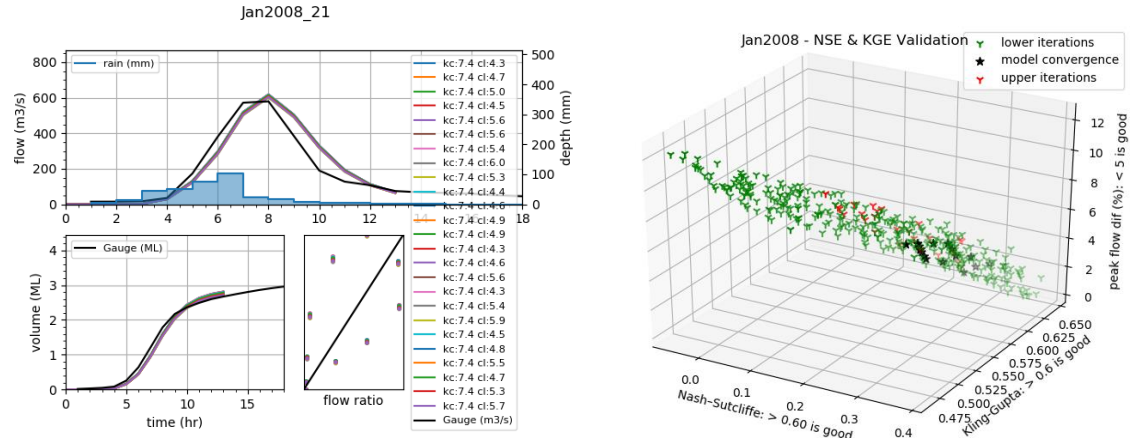

Figure 14. 125006A - Jan 2008 Calibration Results

Table 4. 125009A - Ensemble-learning model results

Event	Kc Range	CL Range	Commentary
Jan 2019	24.1 – 26.2	1.8 – 2.5	Good fit to hydrograph shape and volume-duration. Poor calibration when compared to the resulting performance criterion. Likely due to storm volume, hyetograph shape and/or baseflow.
Mar 2017	24.0 – 26.0	2.7 – 3.7	Poor calibration when compared to the resulting performance criterion. Likely due to storm volume, hyetograph shape and/or baseflow.
Mar 2016	24.0 – 25.9	3.0 – 4.0	Similar to Mar 2017 event
Apr 2018	22.2 – 23.9	2.9 – 3.2	Good fit to hydrograph shape and volume-duration was achieved. Time lag between the pluviograph and gauge was observed that wasn't captured.
Jan 2021	21.1 – 21.7	2.1 – 2.2	Good fit to hydrograph shape and volume-duration. Difference in peak shape likely due to the pluviograph not capturing the catchment average hyetograph shape accurately

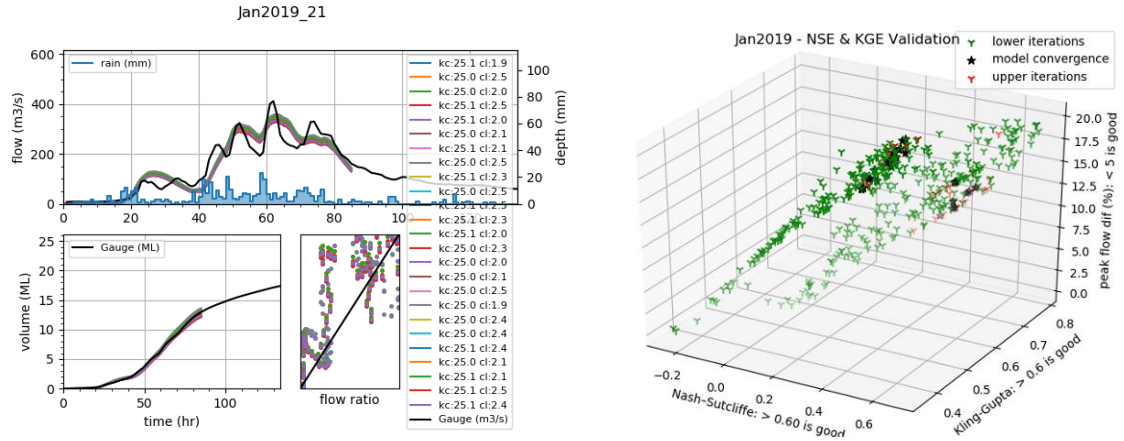

Figure 15. 125009A – Jan 2019 Calibration Results

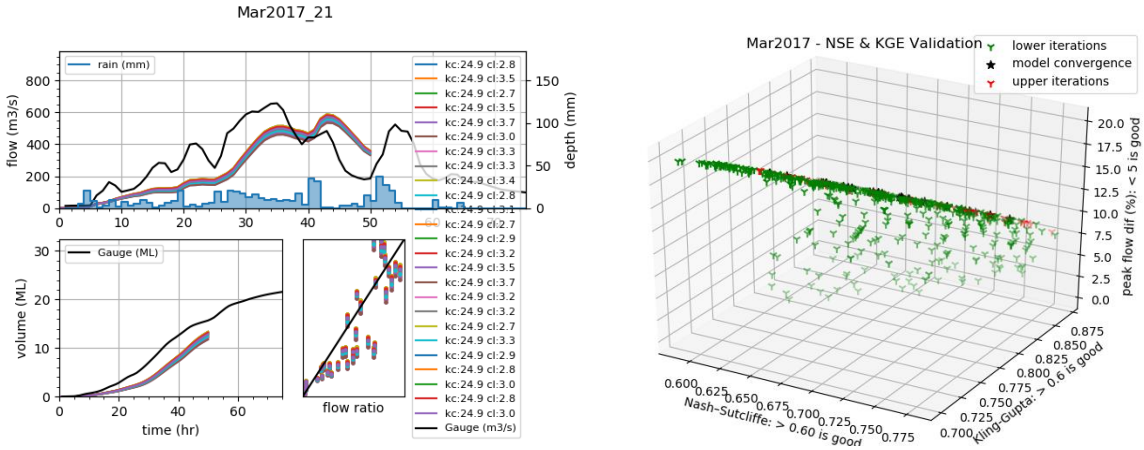

Figure 16. 125009A –Mar 2017 Calibration Results

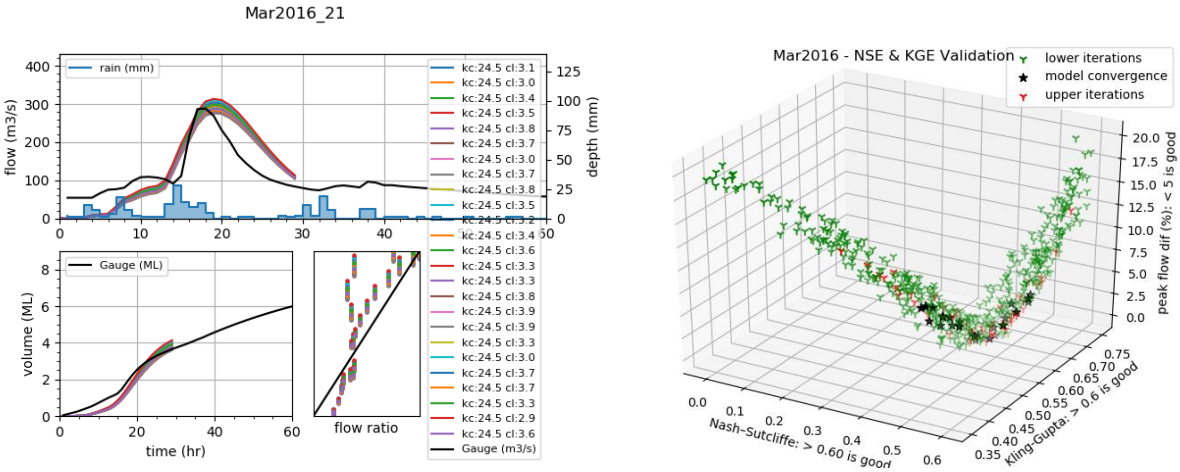

Figure 17. 125009A – Mar 2016 Calibration Results

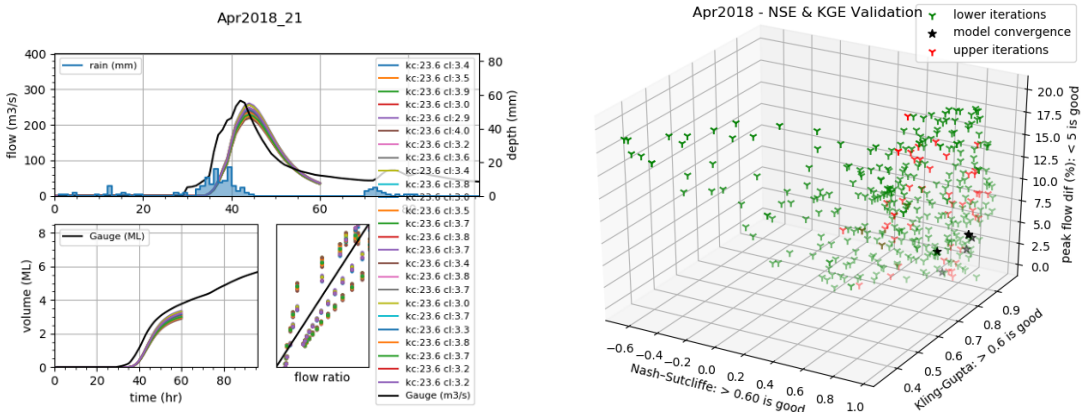

Figure 18. 125009A – Apr 2018 Calibration Results

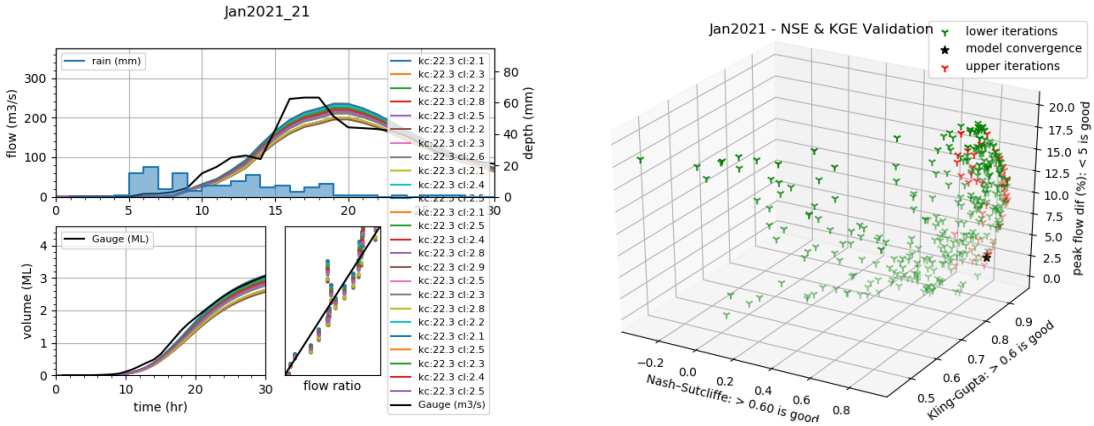

Figure 19. 125009A – Jan 2021 Calibration Results

Table 5. 126003A - Ensemble-learning model results

Event	Kc Range	CL Range	Commentary
Mar 2017	9.8 – 10.4	1.9 – 2.6	ML algorithm was fixed to 30 hrs of simulation time to prevent failure. Good fit to hydrograph and volume-duration suggests the failure is likely due to base flow.
Feb 2008	6.1 – 6.6	1.9 – 2.8	Good observed fit to hydrograph shape and volume-duration. A poor performance criterion result and lost volume over time suggests the need to consider baseflow.
Jan 2013	9.2 – 10.1	1.5 – 2.2	Poor calibration when compared to the resulting performance criterion. Likely due to storm volume, hyetograph shape and/or baseflow.
Jan 2010	5.4 – 5.9	1.5 – 2.1	A good observed fit to hydrograph shape and volume-duration. A poor performance criterion result and lost volume over time suggests the need to consider baseflow.
Mar 2012	-	-	ML algorithm failed during calibration

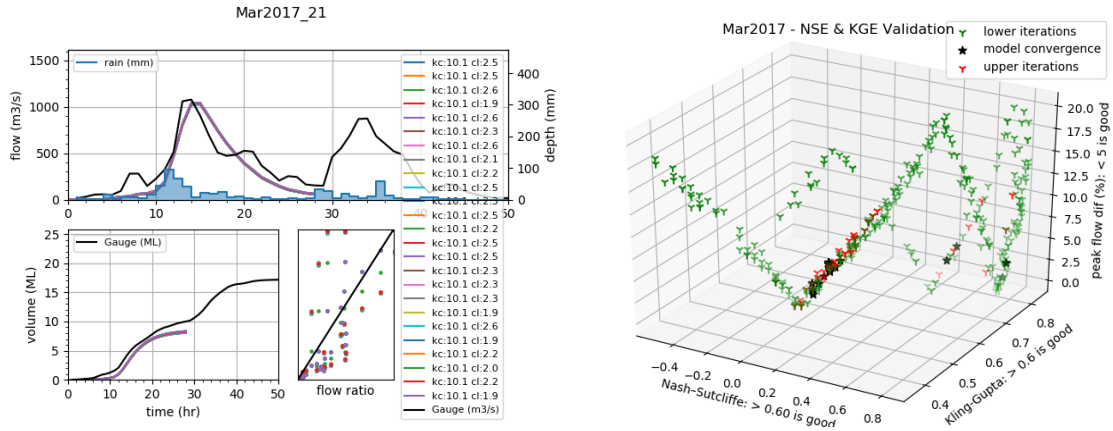

Figure 20. 126003A – Mar 2017 Calibration Results

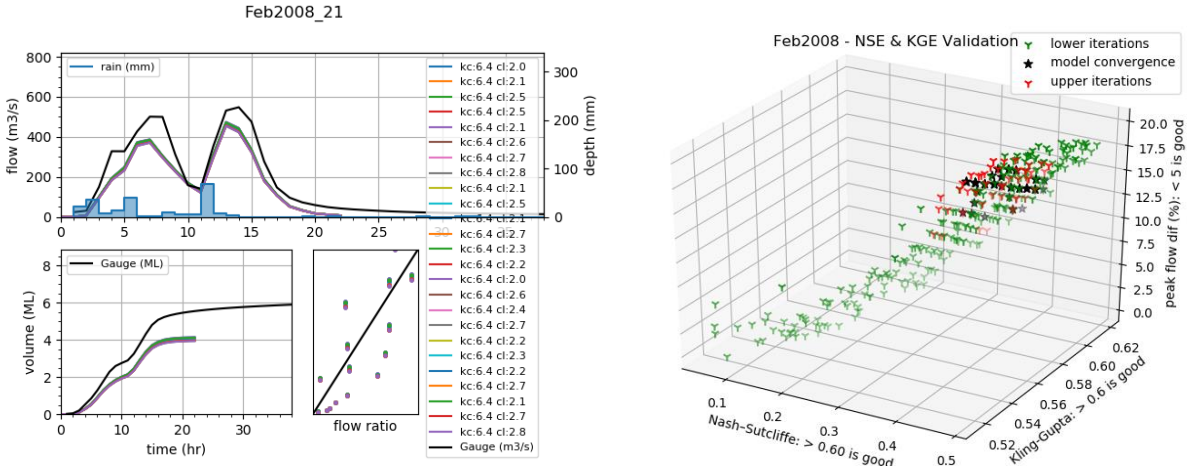

Figure 21. 126003A – Feb 2008 Calibration Results

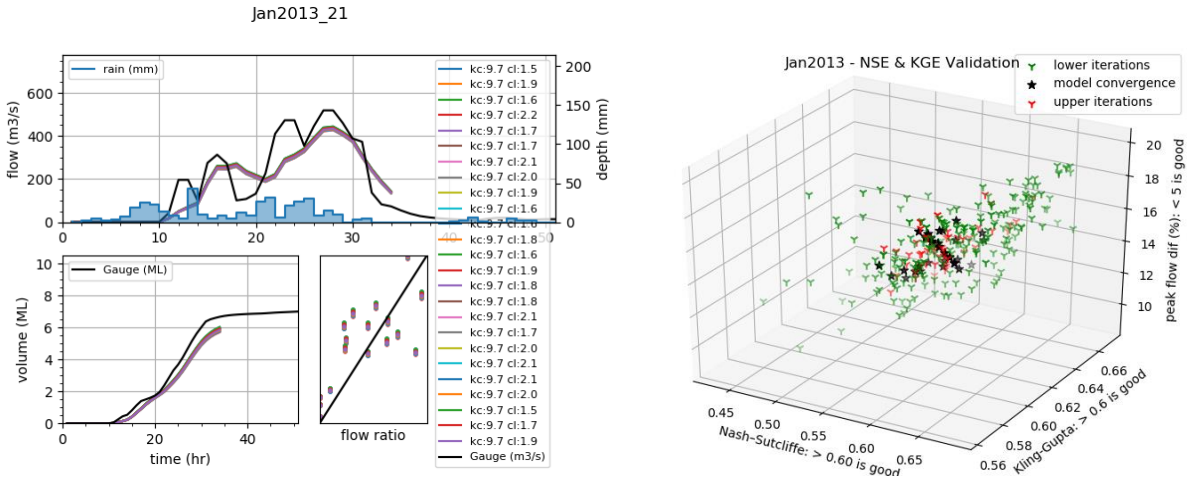

Figure 22. 126003A – Jan 2013 Calibration Results

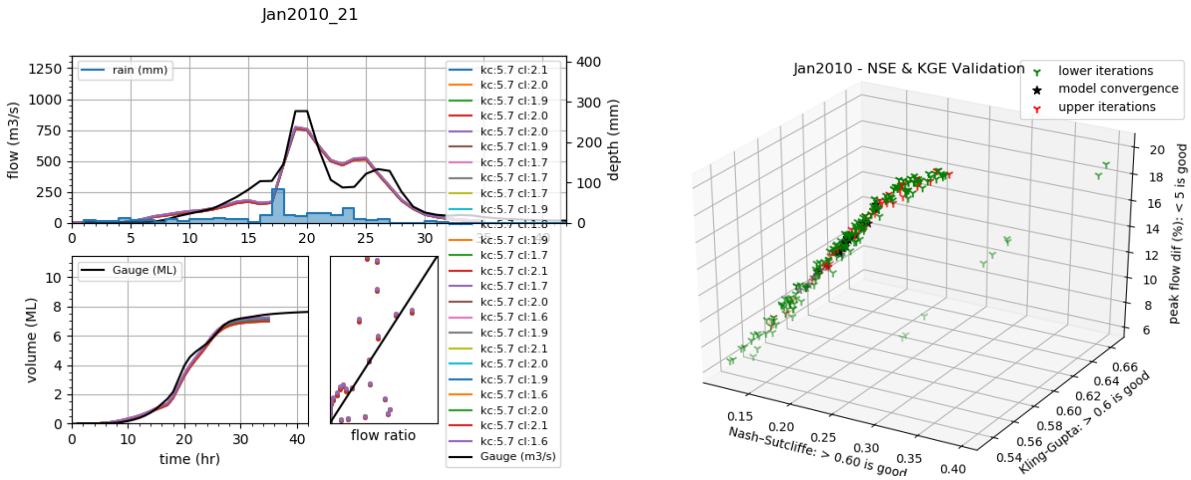

Figure 23. 126003A – Jan 2010 Calibration Results

Table 6. 136301B - Ensemble-learning model results

Event	Kc Range	CL Range	Commentary
Dec 2010	53.2 – 57.8	2.3 – 2.7	Good fit to volume-duration was observed. Hydrograph shape suggests the catchment average hyetograph was not captured accurately.
Jan 2013	47.8 – 52.6	2.2 – 2.9	Close to perfect fit when benchmarked to the performance criterion. Difference in shape towards the hydrograph peak is likely due to hyetograph shape.
Jan 2011	32.2 – 34.4	1.7 – 2.4	Insufficient volume at the tail end of the hydrograph and volume-duration plot, suggests the need to consider baseflow or insufficient hyetograph volume.
-	-	-	Insufficient overlap of pluviograph and gauge records

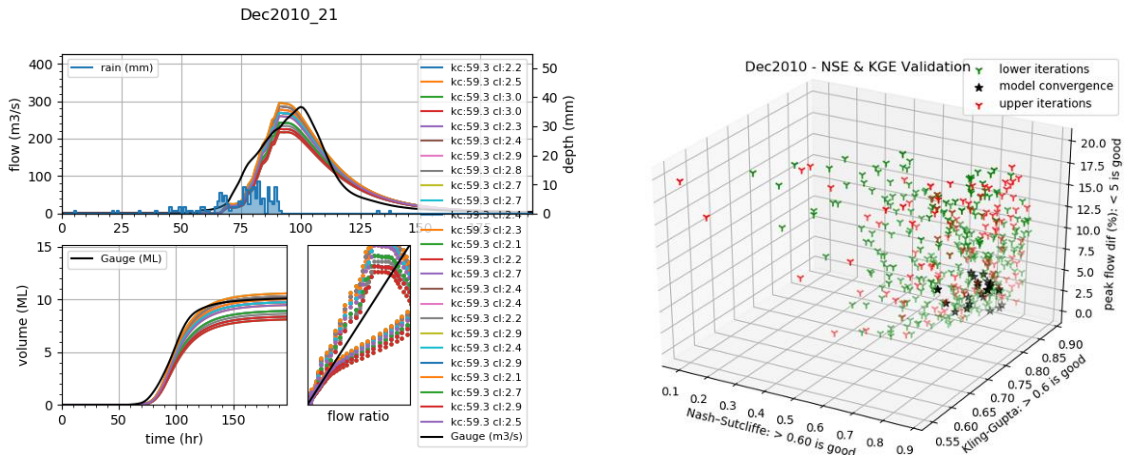

Figure 24. 136301B –Dec 2010 Calibration Results

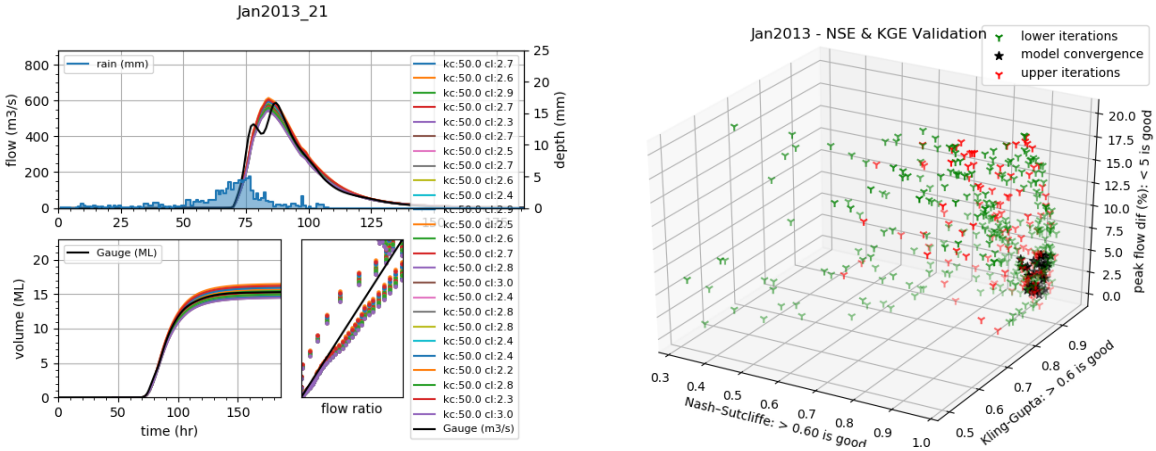

Figure 25. 136301B –Jan 2013 Calibration Results

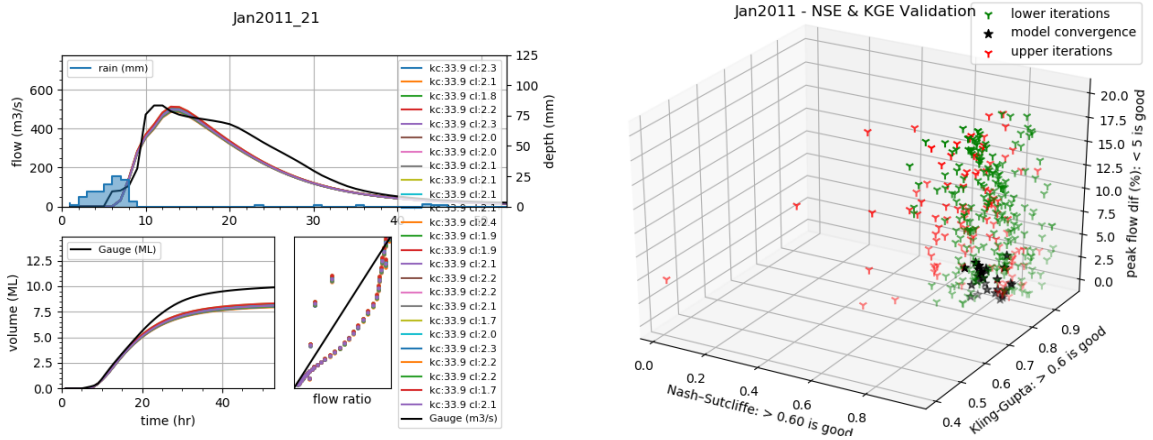

Figure 26. 136301B –Jan 2011 Calibration Results

CONCLUSIONS AND OBSERVATIONS

The ensembled MLP-R ML approach as use in this study was found to perform well when used to estimate RORB input parameters with the following observations:

- Sufficient volume in the model from the storm hyetographs was critical to produce a reliable calibration. Volume-duration should be considered in conjunction with hydrograph shape;
- The hyetograph shape needs to adequately reflected the behaviour of the storm across the catchment (this is particularly observed in the largest catchment (136301B with an area of 500 km^2) where a very good calibration was achieved against the performance criterion;
- In applying spatially varying rainfall, 136301B had better calibrations compared to smaller catchments with single pluviograph rainfall. This reinforces the importance of spatially varying rainfall for larger catchments;
- Where baseflow was observed for calibration events, the machine learning model struggled to converge to a range of parameters or failed in the calibration. This suggests the need for a hydrological model that can also model baseflow when baseflow is observed;

The use of RORB when compared with other hydrological software packages was useful in simplifying the machine learning approach and observing the impacts of input uncertainty in the model calibration. This study recommends that further work be conducted with more complex hydrological software packages to further reduce modelling variables and thereby the associated problem of bias and consequently the risk of underestimation of flood damage.

REFERENCES

Catchment Simulation Solutions, CatchmentSim, NSW, 2021

Dyer, B.G., Nathan, R.J., McMahon, T.A. and O'Neill, I.C. (1995), Prediction Equations for the RORB Parameter kc Based on Catchment Characteristics, *Australian Journal of Water Resources*, 1(1), 29-38

Geoscience Australia, Geoscience Australia, 1 second SRTM Digital Elevation Model (DEM). Bioregional Assessment Source Dataset, 2011.

Gupta HV, Kling H, Yilmaz KK, Martinez GF. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology* 377:80–91.

Nash, J. E.; Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models part I — A discussion of principles". *Journal of Hydrology*. 10 (3): 282–290. Bibcode:1970JHyd...10..282N. doi:10.1016/0022-1694(70)90255-6.

Pearse, M., Jordan, P. and Collins, Y. (2002), A simple method for estimating RORB model parameters for ungauged rural catchments. Instn. Engrs. Australia, 27th *Hydrology and Water Resources Symposium*, CD_ROM, 7 pp

Scikit-learn, Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Weeks, W.D. (1986), Flood estimation by runoff routing - model applications in Queensland. Instn. Engrs. Australia, Civil Engg. Trans., CE28(2), 159-166.

BIOGRAPHY

Kyle Thomson is a Chartered Professional Engineer (CPEng), Registered Professional of Engineer of Queensland (RPEQ) and Industry Engineer. He is based at Water Modelling Solutions with key projects involving flood engineering, drainage, impact assessments, mitigation and flood management strategy, erosion and scour protection, temporary works, linear infrastructure, rehabilitation, CFD and automation. Kyle has delivered work across Australia and overseas using a wide range of software tools and packages.