Water Quality

water e-journal
Online Journal of the Australian Water Association
ISSN 2206-1991
Volume 2 No 2 2017
https://doi.org/10.21139/wej.2017.014

TRANSLATING BIG DATA MODEL OUTPUT TO INFORM POLICY AND DECISION MAKING

Tools and techniques to turn strings of numbers into logical information for technical experts and the community

M van der Sterren, M Griffith, S Manning, P Tate, J Dixon

ABSTRACT

The Hawkesbury-Nepean River system covers approximately 22,000 km^2 and supplies the majority of the drinking water to the city of Sydney. The population of Greater Sydney is expected to grow by approximately 1.5 million people over the next 20 years, with most of the growth occurring in the Hawkesbury-Nepean River catchment. To enable a holistic understanding of the potential impact of this growth on the local waterways, Sydney Water led the development of a catchment-wide hydrodynamic and water quality model for the Hawkesbury-Nepean River system. Sydney Water ensured multi-government stakeholder input resulting in the development of this cutting-edge model for internal decision making and government policy development. Since the finalisation of the calibration and validation of the modelling in 2013, over 130 scenarios have been run requiring the development of new tools and techniques to analyse the large volume of model output. Each scenario run generates approximately 100 GB of output from over 40,000 sites (nodes) in the model. The model has been run with a combination of existing and (possible) future wastewater treatment plants, a variety of treatment technologies, variable wastewater discharge locations, environmental flow options, diffuse runoff control, and future land-use scenarios. Various tools and techniques have been developed in-house to communicate the relative outcomes from the model to key decision makers and stakeholders. This paper explains the analysis methods used to inform further modelling and management decisions.

INTRODUCTION

The Hawkesbury-Nepean River catchment is one of the largest coastal basins in NSW, being approximately 22,000 km^2 (Figure 1). The catchment is planned to be Sydney's next largest urban growth area with the majority of this growth in the South Creek catchment. Increasing urbanisation will not only result in a significant increase in demand for potable water, but will also result in changes in point and diffuse sources of pollution to the Hawkesbury-Nepean River and tributaries.

There are many anthropogenic influences in the catchment resulting in a complex interaction between altered natural river flow, runoff from urban development, discharges from wastewater treatment plants, and provision of reticulated and irrigation water to Greater Sydney. To understand this complex system in more detail from a water quality and environmental flow perspective, Sydney Water led the development of a catchment-wide water quality model. The Hawkesbury-Nepean River and South Creek model (HN model) was prepared by Jacobs Pty Ltd in partnership with BMT WBM Pty Ltd, eWater, University of Western Australia and Yorb Pty Ltd.

Sydney Water ensured multi-government stakeholder input resulting in the development of this cutting-edge model for internal decision making and government policy development. The HN model output is being used by Sydney Water to develop infrastructure solutions for the new growth centres that consider the receiving water environment. The model has also been used to contribute to the NSW Government Environmental Flows project to assess the impact of environmental flows from Warragamba Dam on the health and quality of the river. A number of government departments were consulted in the design of the HN model.

The HN model has been developed to look at key water quality management issues within the river. These issues include water pollution, algae, aquatic weeds, flows rates, and flow distributions. The HN model has the ability to simulate hydrological, hydraulic and biogeochemical processes to examine the water quality benefits (or impacts) resulting from the different management strategies over broad spatial and temporal scales. Over 130 scenarios have been run through the model, testing a combination of population growth, wastewater treatment quality and discharge location, land-use, diffuse source management, and environmental flows.

This paper focusses on the analysis of the output generated from the catchment and 3D hydrodynamic and water quality model to inform decisions. Each scenario run generates approximately 100 GB of model output from over 40,000 sites (nodes) in NetCDF files. While any of these 40,000 nodes can be interrogated, 52 sites (nodes) were identified by Sydney Water and other government stakeholders as key locations for detailed analysis. This reduced each file size to approximately 2 GB. However, with over 130 scenarios, this still results in a large data set requiring manipulation and analysis. The sheer volume of results fall under the definition of Big Data. Coombes and Barry (2014) defined big data analysis as *'the term for a collection of large and complex datasets that are difficult to process or understand using traditional database management tools or data processing applications'.* Due to the complexity of the underlying assumptions and scenario settings, interpreting the modelled results can be difficult using traditional statistics in conventional analysis. As such, Sydney Water has developed a number of analysis techniques to communicate the possible outcomes of various management options to stakeholders and Sydney Water's regulator (NSW Environment Protection Authority, (EPA)).

This paper highlights the analysis used and lessons learned.

THE HAWKESBURY-NEPEAN RIVER AND SOUTH CREEK MODEL SETUP

The HN model simulates the flow and quality in the river system from Warragamba Dam on the Warragamba River, and Pheasants Nest and Broughton Pass weirs downstream of the Upper Nepean dams, to the ocean, a distance of 260 kilometres (Figure 1).

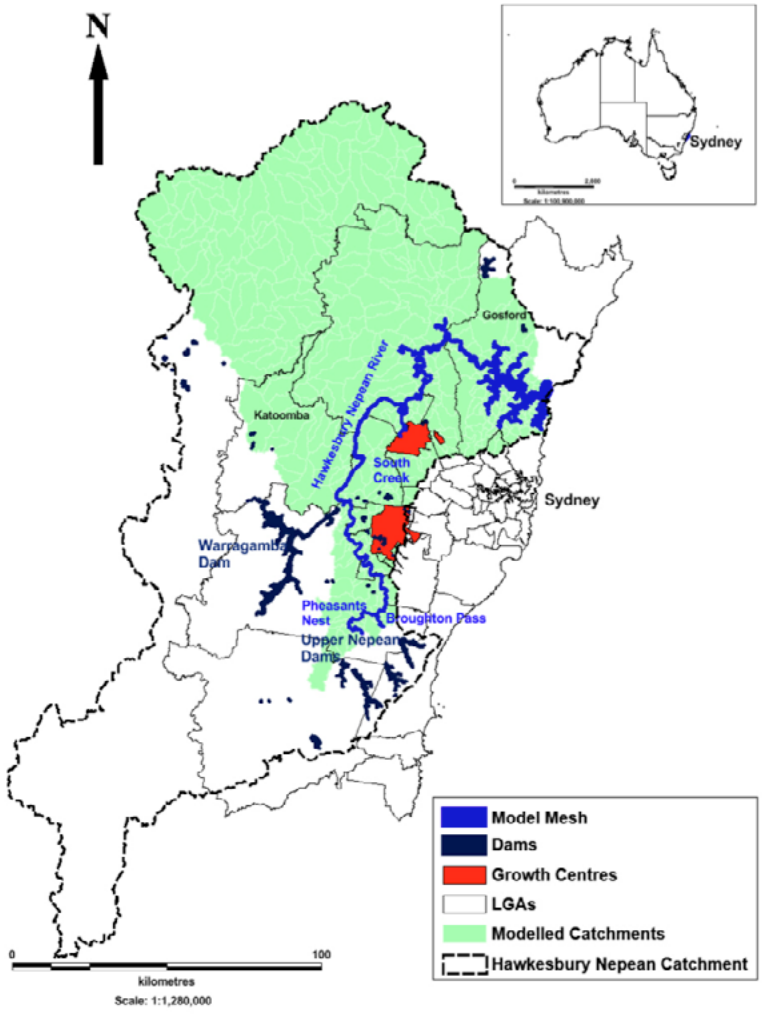

Figure 1. Hawkesbury-Nepean River system and its catchment

The HN model comprises four linked models:

- Source catchment model
- TUFLOW FV (Three dimensional, Unsteady FLOW, Finite Volume) hydrodynamic model
- Aquatic EcoDynamics (AED) water quality model
- EcoModeller macrophyte model

A conceptual diagram of how the models work together is shown in Figure 2 (Griffith and Tate 2014).

The Source catchment model is used to generate daily flows into the river system downstream of the dams. Flows from the dams were provided by Water NSW. The Source model also incorporates extractions from the river by irrigation (provided by Department of Primary Industries, Water) and flows from wastewater treatment plants (WWTPs). The TUFLOW FV hydrodynamic model uses these flows and simulates the hydrodynamics of the river in three dimensions taking into account tidal levels, flows, salinity, wind and temperature varying over time to produce velocity, depth, salinity and temperature across the river system. The results from TUFLOW FV are incorporated into the AED water quality model to calculate concentrations of nutrients, algae and bacteria in the river system over time. The water quality and quantity outputs are imported into EcoModeller for ecological modelling. EcoModeller uses rating curves and different time scales to generate a relative cover score for a particular macrophyte species, in this case *Egeria densa.*

The Source catchment model was calibrated and validated using results from 1 July 1998 to 31 December 2011. For flow calibration and validation, gauged flow data was randomly separated into calibration years and validation years (SKM, 2014a). The TUFLOW FV model was calibrated for 2006 data and validated using 2007 data. This period had the largest amount of data available in terms of flow and water quality (SKM, 2014a).

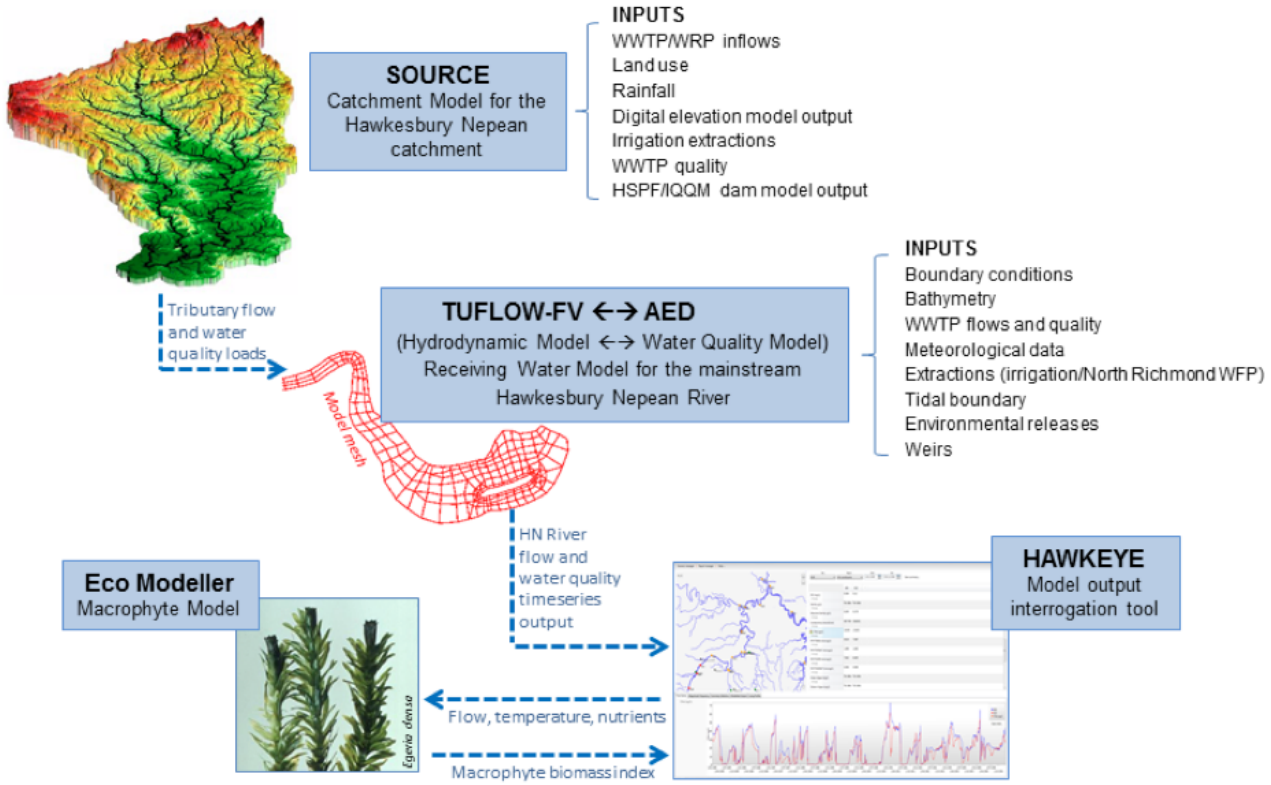

Figure 2. Hawkesbury-Nepean River and South Creek model schematic (Griffith and Tate 2014)

SCENARIO DEVELOPMENT

The HN model has been built for the express purpose of providing guidance on the likely quantitative differences in water quality and quantity when contrasting different catchment and environmental flow, wastewater and land use scenarios over time. Overall differences in the statistical properties of flow and constituent concentrations between scenarios can be inferred by comparing scenarios. This includes differences between mean values, or differences between values that may be exceeded for a given proportion of time. Moreover, while the model has been developed to simulate flow and constituent concentrations at specific locations, the real benefit of the model is in the assessment of the overall outcomes of a particular suite of management actions across the spatial and temporal domain encompassed by the model, compared to an alternative suite of actions or a "do nothing" scenario (SKM 2014b).

The first 100 water quality scenarios were developed as part of an inter-agency working group. The scenarios included:

- Environmental flow release policies, represented as changes in the input time series for flow and water quality from Warragamba Dam to the TUFLOW FV/ AED model.
- Changes to WWTP discharges, represented as daily time series of concentrations from WWTPs into either the catchment model or the TUFLOW FV/ AED model depending on the actual or possible location of the discharge.
- Population growth as represented by changes in land use in the catchment model.
- Implementation of water sensitive urban design in "green field" or new urban areas to limit the loads of sediment and nutrients generated from these areas. These are represented as changes in generation rates for constituents from the new urban land use in the catchment model.
- Assimilation of nutrients in South Creek facilitated by decay functions within the catchment model.

Since completion of the first 100 scenarios, Sydney Water has continued to develop new scenarios for various projects. For each project, it is essential that the objective is clearly defined. This is particularly important as the HN model takes approximately three weeks to run, plus set up time. Key questions to address include:

- What are we investigating? If it is more than one thing, e.g. environmental flows and urban development, multiple compounding scenarios will need to be run and analysed.
- What are the constituents of interest? In some cases, only some of the constituents need to be analysed.
- What are the changes to the input parameters? This can include WWTP treatment capacities or transfers of WWTP flows to different discharge areas.

A detailed project plan needs to be agreed with the stakeholder(s) to ensure that the assumptions are clearly understood. The project plan also needs to specify data analysis and presentation techniques. This process is essential to successfully communicate the analysed model outcomes to the various stakeholders at the appropriate level of detail. A variety of techniques have been developed to maximise the information extracted from the model output. These include:

- Comparison of the modelled constituent concentration to objective values, e.g. Healthy River Commission objectives for the Hawkesbury-Nepean catchment (HRC 1998)
- Comparison of the average constituent concentration between sites and scenarios
- Flow based analysis
- Load analysis
- Whole of catchment animations.

All techniques discussed in this paper have been programmed in MATLAB and use either the NetCDF outputs from TUFLOW FV or the excel outputs from Source, EcoModeller or TUFLOW FV.

SCENARIO ANALYSIS

Dry and Wet Weather Comparisons

The HN scenario model is based on the ten year climate sequence between 1985 and 1994. The 1985-94 period contains a mixture of wet, dry and average weather years. As such, the model output can be extracted and analysed from particular weather conditions if required. Analysing the dry weather model output is of particular importance for Sydney Water to understand potential impacts on water quality during extended dry weather, while land managers may be more concerned with wet weather when diffuse runoff dominates the receiving water quality.

Table 1. HRC (1998) trigger values for the Hawkesbury-Nepean River system

Type of catchment	Mixed use rural areas and sandstone plateau	Urban areas main stream	Urban areas tributary stream	Estuarine area
Total phosphorus (mg/L)	0.035	0.03	0.05	0.03
Total nitrogen (mg/L)	0.7	0.5	1	0.4
Chlorophyll-a (µg/L)	7	10-15	~20	7

All the plots presented below can be generated as 'all weather' (i.e. using the full 10-year data series), 'dry weather' or 'wet weather'.

Comparison of Modelled Results to Objective Values

The Australian and New Zealand Environment Conservation Council (ANZECC) has developed a framework for defining, assessing and protecting water quality and ecosystem health, known as the Guidelines for Fresh and Marine Water Quality (ANZECC 2000). These have been adopted as the standard for NSW. The guidelines promote a risk based approach that moves away from universal numeric guideline values in favour of site specific studies. Level one of the risk approach does, however, consist of numeric default trigger values that are deliberately conservative and are intended to be applied when there is no pre-existing knowledge of the subject system. The Hawkesbury-Nepean River system has an extensive dataset including long term and ongoing monitoring of the river. This information was used by the Healthy Rivers Commission (HRC) to complete the next level of the ANZECC framework and set water quality objectives specifically for the Hawkesbury-Nepean River. The Commission published the results of its inquiry into the Hawkesbury-Nepean in 1998 which included nutrient objective levels for five separate zones in the river (HRC 1998). In some cases, the objectives are more stringent than the ANZECC trigger values; in other cases they are more tolerant.

One of the techniques used to analyse the HN model output is based on the Healthy Rivers Commission's objective values (Table 1). Performance is indicated by calculating the proportion of results for a particular constituent/variable that falls within the objective value, expressed as a percentage (Figure 3, right hand side scale, line graph).

Comparison of the output to the HRC objectives is used as a first pass analysis to identify if there is a change in the number of results (model output) outside the objective values. From the example shown in Figure 3, it can be seen that scenario C (purple line) consistently performs the best with respect to the HRC guideline although there is deterioration at site 2 where it is comparable to the other scenarios. Scenario D (light blue line) also has relatively good performance with respect to the HRC objectives.

Average Constituent Concentration

The average is a way of aggregating the data and assessing the relative merits of a variables distribution at a given site under given climate conditions. It allows for a simple comparison between sites and scenarios (Figure 3, left hand scale, bar graph) for a particular constituent. The average is used as a first pass analysis. It takes into account the distribution of the constituents but does not distinguish between different flows.

From the example shown in Figure 3, scenarios A and B (dark blue and green bars) have average modelled concentrations consistently higher and occasionally double that from scenarios C and D (purple and light blue bars). Scenario E (bright blue bar) is more variable. Depending on the importance of the constituent and sensitivity of the particular site, this may guide the preference for the management practices associated with scenarios C and D over scenarios A, B and E. The example provided is very simple and represents only one constituent. Realistically, analysis and interpretation would need to include an array of constituents, locations and scenarios. This analysis may also prompt new scenarios which investigate a modified version of the initial scenarios.

When comparing the two indicators, i.e. the HRC objectives and the average, it is important to understand that the average is a continuum whereas the HRC counts the number of records above and below the relevant objective value. Therefore, for a given scenario the average may change, but not compliance with the HRC objective. Conversely, a small change in the average may be accompanied by a large change in HRC compliance.

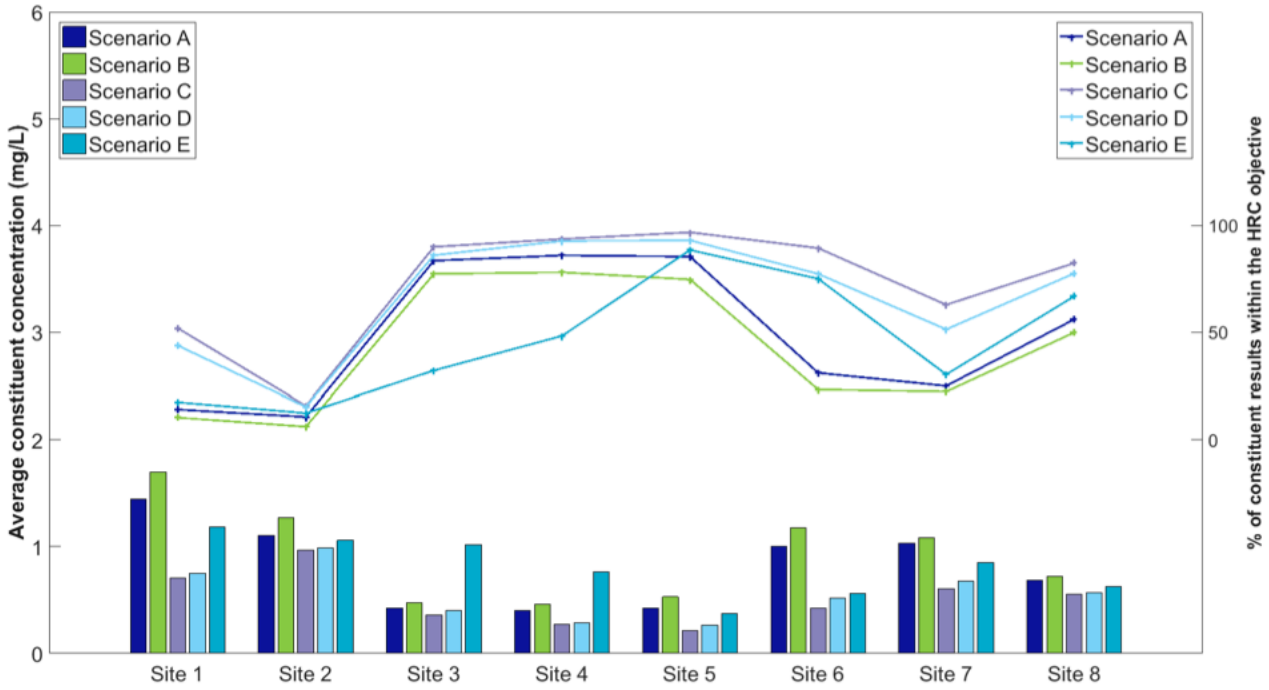

Figure 3. Percent within the HRC objectives and average comparison (from upstream-site 1 to downstream-site 8)

The technique of using both the average and the HRC objectives comparison provides additional information to enable decision makers to assess the benefits, or otherwise, of different management options. The findings from the modelling would then feed into the larger decision making process that considers both social and economic factors, as well as the environment.

Flow and Load Analysis

The methods described above allow for a big picture view of the output. They provide a first pass analysis, which in some cases may be all that is required, depending on the objective of the modelling and analysis. It may also trigger the need for more detailed analysis. Analysing the model output in flow classes is a good way to gain refined information from the scenario results. Of particular interest to Sydney Water is the water quality during a moderate to low flow regime. As such the results were separated and analysed in flow categories. The criteria used are presented in Table 2.

The flow categories were defined using the baseline scenario for a particular site or stretch of river. The base scenario flows were analysed using conventional statistics and flow rates assigned to each flow category. The flow from the comparison scenario was then used to define the flow category for each flow type. These flow divisions assist in the identification of changes in flow regime and changes in water quality as a result of the different management options.

The modelled water quality can be analysed using flow weighted total loads presented as bar plots (Figure 4) or flow weighted concentrations presented as box-whisker plots (Figure 5).

Table 2. Flow categories adopted for analysis

Flow categories	Percentile flow
Very low flows	≤5%
Low flows	>5 and ≤20%
Moderate flows	>20 and ≤70%
Freshes	>70 and ≤90%
Floods	>90 and ≤99%
Extremes	>99%

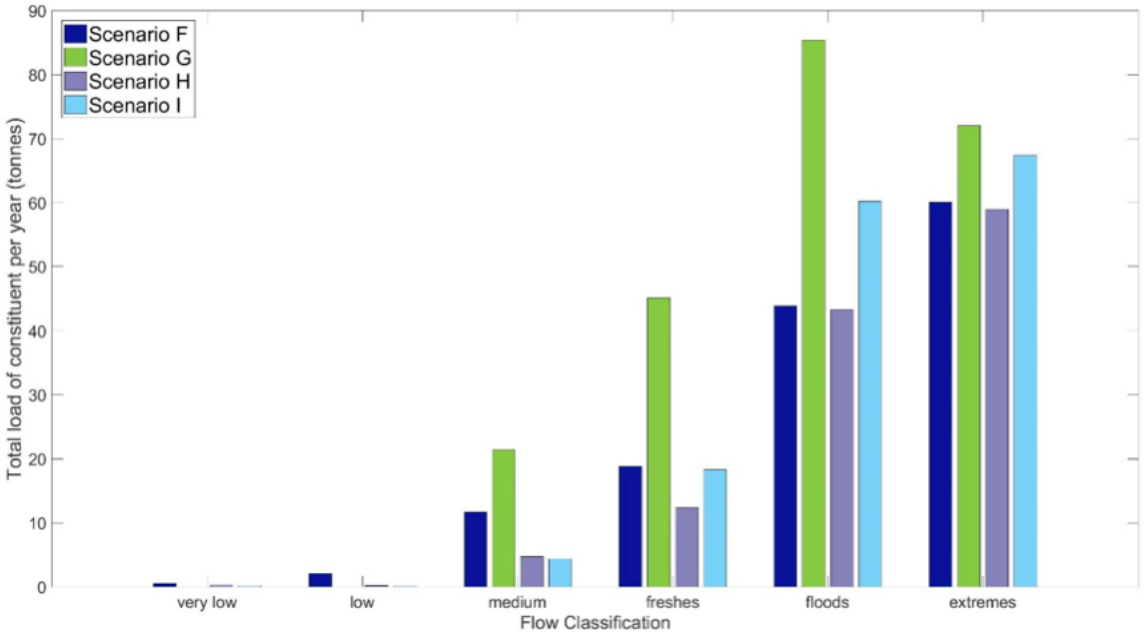

Figure 4. Total load discharged as defined through flow categories

The example shown in Figure 4 shows the loads of a constituent from a tributary/sub catchment entering the mainstream river under different management options and flow conditions. It provides a useful indication of the loads and performance of the scenarios under certain flow conditions.

Box and whisker plots require the data to be approximately normally distributed and each dataset to be independent (Milton and Arnold, 2003). It is assumed that breaking up the flows into different categories results in each category consisting of paired, independent datasets, therefore allowing the use of box-whisker plots.

Presenting the modelled concentration data as flow weighted box and whisker plots for a particular site provides further insight into the scenarios (Figure 5). This is of particular importance at very low and low flows when higher concentrations are more likely to impact on water quality than much lower concentrations.

This analysis has proven useful to look at the changes occurring as a result of development and increased flows from wastewater treatment plants. The impact of higher flows with lower concentrations is not visible from a non-flow separated graph, but becomes clear using a flow division. This allows for an improved understanding of the potential impact of the proposed solution by policy developers and decision makers. Furthermore it facilitates discussion on the importance of understanding both concentration and load in managing water quality.

ANIMATIONS

Animations of the modelled outputs were created using MATLAB software. They provide a high level visualisation of how concentrations can vary for different scenarios over a given time across the whole, or part, of the model domain. Any time period can be chosen depending on the objective, i.e. is the time period of interest related to a widespread storm event or an extended dry period? The constituent concentration is represented by colour and height of the river in the animation. Each animation is created by extracting approximately 20 GB of NetCDF model output previously generated by the HN model. The animation uses the output from all the 40,000 plus sites (nodes), as opposed to just the 52 key sites. Users can change the zoom, angle and scale of any axes to change the view of the Hawkesbury Nepean River as well as selecting only certain sections of the river for display. The animations were found to be a useful technique to present the model results and outcomes to management and the wider community.

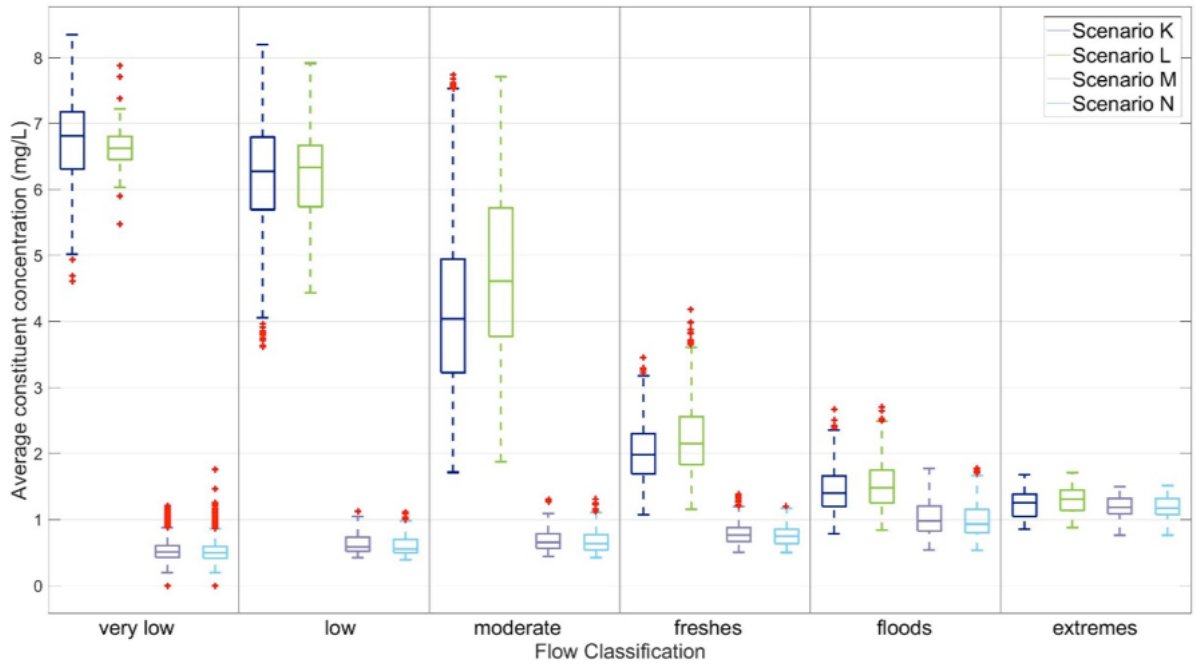

Figure 5. Constituent concentration in a waterway separated into flow classes

APPLICATION OF THE MODEL OUTPUT TO MANAGEMENT DECISIONS

The model output is currently being used by Sydney Water to understand the water quality impacts that may occur in response to the different wastewater treatment plant options being considered to service population growth in the Hawkesbury Nepean catchment. It is also being used to provide scientific evidence that will contribute to the licence renewal discussions with the EPA.

The population growth planned for the Hawkesbury-Nepean catchment brings many changes that have the potential to impact local waterways, more than wastewater treatment and their associated discharges. The HN model has the functionality to test a variety of catchment-wide initiatives and provide an improved understanding of the role of diffuse source pollution. As such it is anticipated that the model will prove highly beneficial for all NSW government departments in the future.

CONCLUSION

The ability to handle and process big data to inform policy and management decisions, is becoming increasingly important. With rapid advances in technology combined with increased computing power, this is now achievable. Multiple lines of evidence are used to develop policy and make decisions, of which modelling forms one component. To ensure that complex modelled outcomes are understood, the development and use of innovative analysis techniques is key to communicate with management, stakeholders and the community. This paper presents different techniques used to analyse the water quality model outputs from the Hawkesbury-Nepean River and South Creek model.

The development of new scenarios and analysis techniques resulted in a number of lessons learned. Defining the objective and ensuring that all assumptions were clearly understood was found to be critical when establishing new scenarios and communicating the findings to stakeholders.

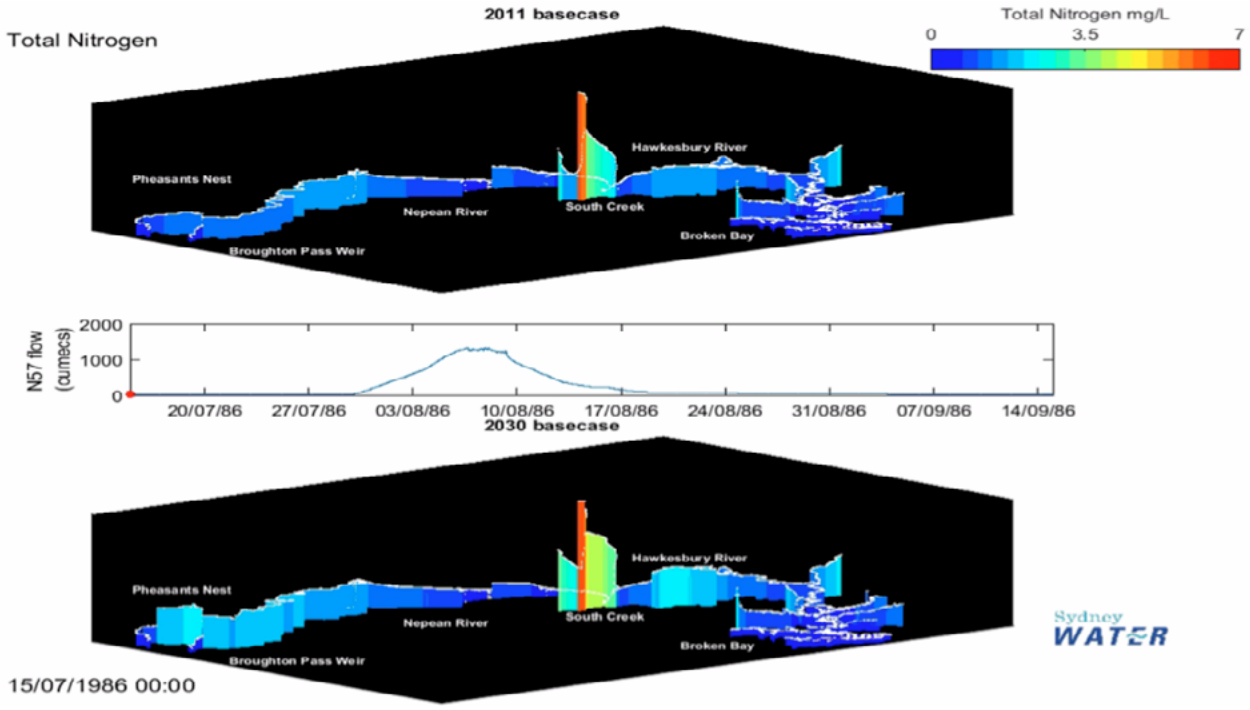

Figure 6. Animation comparing two scenarios generated from the HN model

It was also found to be beneficial to keep the initial analysis simple, presenting the average constituent concentration in conjunction with a comparison to relevant guidelines or objectives. Depending on the objective, this may prompt more detailed analysis using flow separated statistics. Separating the load or concentration by flow classes provided an improved understanding of how increased flows with lower concentrations or decreased flows with higher concentrations, influence the water quality within the river. The use of animations proved to be an effective technique to communicate the overall findings to management and the community, however it is not suitable for detailed analysis. Work is continuing to develop more robust techniques to analyse, compare and contrast the different scenarios to enable decision making by various NSW Government agencies within the Hawkesbury-Nepean River catchment.

ACKNOWLEDGMENTS

Sydney Water would like to acknowledge Jacobs Pty Ltd (previously Sinclair Knight Merz), BMT WBM, eWater, Yorb Pty Ltd, and UWA for the development of the model and running of over one hundred scenarios, the CSIRO for the review of the model, and Natalie Marshall (Sydney Water) for running additional scenarios and contributing to the analysis.

Extensive data sets were provided by DPI Water (previously NSW Office of Water), Office of Environment and Heritage, Water NSW (previously Sydney Catchment Authority), Manly Hydraulics Laboratory, Land and Property Information, Bureau of Meteorology, Hornsby Shire Council, Penrith City Council, The Hills Shire Council, Blacktown Council and Camden Council.

THE AUTHORS

Dr. Marlène van der Sterren, is a Principal Modeller in the Asset Knowledge team at Sydney Water. She is the principal modeller responsible for the operation and maintenance of the Hawkesbury Nepean River and South Creek model. Marlene is a Chartered Professional Engineer in the Civil and Environmental colleges of Engineers Australia (CPEng, MIEAust) and a Certified Professional in Erosion and Sediment Control.
marlene.vandersterren@sydneywater.com.au

Merran Griffith, is the Principal Advisor Waterway Health in the Corporate Strategy team at Sydney Water. Merran project managed the build of the Hawkesbury Nepean River and South Creek model. She has 24 years experience in the water industry with a strong focus on water quality and environmental monitoring. Merran holds a Bachelor of Applied Science in Environmental Biology and a Masters in Environmental Management.
merran.griffith@sydneywater.com.au

Scott Manning, is an Analyst in the Corporate Strategy team at Sydney Water. He is responsible for developing tools for statistical analysis and visualisation of model outputs. Scott has a Bachelor of Civil (Environmental) Engineering and Science.
scott.manning@sydneywater.com.au

Dr Peter Tate, was an Analytics Strategist in the Corporate Strategy team at Sydney Water. He is currently the Director of WQ Data Pty Ltd. Pete has over 35 years' experience in the wastewater industry, focusing on modelling and monitoring of environmental impacts. He has a PhD in Applied Mathematics, focusing on the rise and dilution of buoyant jets and their behavior in an internal wave field.

Dr Jonathan Dixon, is a Principal Analyst in the Corporate Strategy team at Sydney Water. He is currently developing receiving water models for Sydney Water. Jonathan holds a doctorate in Pure Mathematics.
jonathan.dixon@sydneywater.com.au

REFERENCES

ANZECC (2000) Australian and New Zealand Guidelines for Fresh and Marine water quality. Australian and New Zealand Environment and Conservation Council and Agriculture and Resource Management Council of Australia and New Zealand, Canberra, 1-103.

Coombes, P. J. and Barry, M. E. (2014), A systems framework of big data driving policy making – Melbourne's water future, OzWater14 Conference, Australian Water Association, Brisbane.

Griffith, M. and Tate, P. (2014) Hawkesbury Nepean River and South Creek model: A review of a new tool to inform management decisions in the Hawkesbury Nepean catchment. Australian Water Association Water Journal 41(8): 51-57.

Healthy Rivers Commission (HRC) (1998) Independent Inquiry into the Hawkesbury-Nepean River System. Final Report August 1998. Healthy Rivers Commission of NSW.

Milton, J. S. and Arnold, J. C. (2003), Introduction to probability and statistics – Principles and applications for Engineering and Computing Sciences, 4th Edition, McGraw Hill, Sydney.

Sinclair Knight & Merz (2014a), Water quality modelling of the Hawkesbury-Nepean River System – Hawkesbury-Nepean River and South Creek Model – Final Calibration Report. Produced for Sydney Water.

Sinclair Knight & Merz (2014b), Water quality modelling of the Hawkesbury-Nepean River System – Hawkesbury-Nepean River and South Creek Model – Summary Report. Produced for Sydney Water.