22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

Concepts, Philosophy and Methods for Development of a General Linear Statistical model for River Water Quality

Freeman J. Cook [a], **U. Khan** [b], **R. Laugesen,**[b] **G. Amirthanathan**[b], **N. K. Tuteja**[b] **and M.A. Bari** [b]

[a] *Freeman Cook and Associates PTY LTD; The University of Queensland School of Food and Agriculture; Griffith University School of Environment*
[b] *Water Forecasting Services, Bureau of Meteorology*

Email: freeman@freemancook.com.au

Abstract: The modelling of water quality is of considerable importance understanding how to intervene or manage the water quality output from catchments. Modelling methodologies exist from bottom up mechanistic models, through hybrid models to statistical models. Here we will describe a statistical modelling methodology which was developed for forecasting water quality in the Great Barrier Reef catchments using a general linear model in terms of the concepts, philosophy and methods.

The conceptual model considers the pathways that exist for constituents (solutes and particulates) in the streamflow to be transported from the soil in the catchment to the stream (Fig. 1). These can be grouped into those flowing over the surface of the soil where there is an interchange between the soil and the surface soil, including exfiltration and flow coming into the stream via groundwater flow mainly as baseflow.

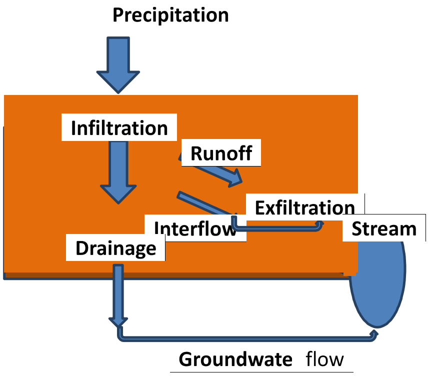

Figure 1. Schematic diagram of components of flow from a catchment to stream

The philosophy of this model is to create the most efficacious model using the least complex modelling framework with easily available data. Streamflow components of: the hourly streamflow (Q (m^3 s^{-1})); hourly baseflow (q (m^3 s^{-1})); the sum of Q minus the long-term mean flow (Q_m (m^3 s^{-1})) were chosen to represent the flow components. Q_m represents a measure of the catchment condition prior to the flow, with negative values representing dry and positive values wet catchment conditions. The differentials of these flow components with time (t (s)) were calculated. The sign of the differentials represents the rising or falling limb of the hydrograph and the magnitude the rate of change. Integrals of Q, q and Q_m were calculated for different lengths of times prior to the measurement time. Transforms of all these components with a power of 0.2 and 2 were calculated. Finally, two orthogonal periodic time-based components sin($2\pi t/86400P$) and cos($2\pi t/86400P$) with P = 365 day were calculated. These covariates are first filtered in a univariate way with the seven constituents: Total Suspended Solids (TSS), Particulate Nitrogen (PN), Dissolved Inorganic Nitrogen (DIN), Dissolved Organic Nitrogen (DON), Particulate Phosphorus (PP), Dissolved Inorganic Phosphorus (DIP) and Dissolved Organic Phosphorus (DOP). With the transforms, this results in 77 possible univariate relationships. These are filtered so only those with an $r^2 \geq 0.1$ are selected subsequent analysis using a multivariate additive general linear model. The fitting process also allows the constituent to be transformed and the best model selected with a maximum of five covariates. This process is semi-automated and can result in millions of possible models which reduced in a selection processes to give the model that best fits measured constituent data. The data was split so that validation could be attempted. This model can then be used along with the streamflow data to estimate the constituents in hindcast and forecast mode and uncertainty estimated.

Keywords: *Water quality, statistical model*

1. INTRODUCTION

Modelling of water quality is difficult because, unlike water quantity where the driver is potential energy or pressure, the transport is driven by the velocity of the flow. Models to estimate water quality at catchment level can vary from processes based models such as SWAT (Arnold et al., 1985), MIKE-SHE (Abbott et al., 1988) to semi-distributed models such as SOURCE (Cook et al., 2009) to statistical models (Kuhnert et al., 2012; Robson and Dourdet, 2015). The models of Kuhnert et al. (2012) and Robson and Dourdet (2015) are based on the original model by Wang et al. (2011). The problem with the distributed and semi-distributed models is to arrange large data sets to populate the models and the computational effort required. Whereas the statistical models are reliant on: i) the quality of both water quantity and quality data, ii) the breadth of the data from low to high flows and low to high concentrations, and iii) the total length of the data is sufficient and typical of the catchment of interest.

The eReefs project (http://www.ereefs.org.au/), which is a collaboration between number of agencies to develop an operational real-time forecasting marine model of water quality for the Great Barrier Reef Lagoon (GBR), requires input of water quality from catchment models to the estuarine and lagoon hydrodynamic models. Previously (Cook, 2013) reviewed modelling approaches for modelling water quality in the eReefs project and concluded that a statistical approach is advisable. Statistical models have been widely used for water quality modelling based on the well-known rating curve method models (Miller, 1951; Cohn et al., 1992). Here we will present the concepts, philosophy and methodology that was developed to produce a statistical water quality model that allows forecasting of the water quality from catchments discharging into the Great Barrier Reef Lagoon. We will illustrate the model with some selected model output.

1.1. Problem definition

The problem is to provide a robust model for water quality from end of system gauging stations on selected rivers in the GBR, that allow for forecasting using forecast rainfall and hence streamflow and can give uncertainty bounds on the estimated concentration and load. The constituents that were selected for modelling were: Total Suspended Solids (TSS), Particulate Nitrogen (PN), Dissolved Inorganic Nitrogen (DIN), Dissolved Organic Nitrogen (DON), Particulate Phosphorus (PP), Dissolved Inorganic Phosphorus (DIP) and Dissolved Organic Phosphorus (DOP), as these are required important constituents for the eReefs marine models. What we sought was a correlation between covariates based on gauged streamflow and measured point samples of the constituents. The constituent data used is that from the GBR loads monitoring program and spanned the time from 2006 to 2014 but for some sites the starting date was later than 2006. The hourly streamflow data was obtained for the site at which the constituents were measured or at the nearest site to this and the time of measurement shifted according to the rules in Turner et al. (2012).

2. CONCEPTUAL MODEL PHILOSOPHY AND COVARIATES

The conceptual model is shown in Figure 1 and considers the pathways by which constituents (solutes and particulates) are transported from the catchment to the river. There are pathways missing from this diagram in the form of the hyporheic exchange processes, streambank erosion, sedimentation and resuspension, biological transformations and vapour loses and deposition. These processes are accounted for in the covariates we use in this statistical modelling framework.

The covariates are based on the streamflow (Q (m^3 s^{-1})), baseflow (q (m^3 s^{-1})) and a covariate we derive to account for the catchment condition, Q_m (m^3 s^{-1}) given by:

$$Q_m = \sum_{t=t_o}^{t_n} \left(Q_t - \bar{Q} \right) \quad (1)$$

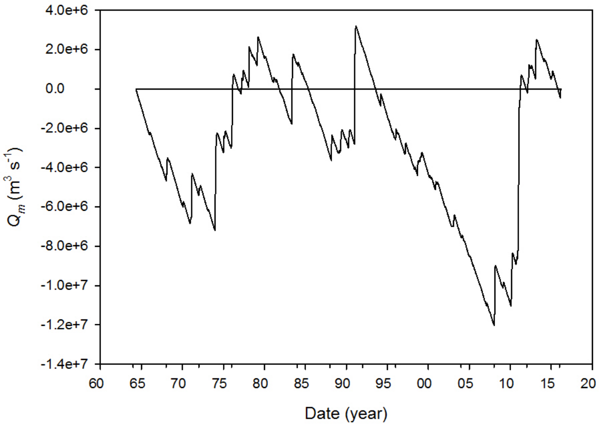

Figure 2. Q_m versus time for the Fitzroy catchment from data of streamflow at The Gap. Note three distinct long dry periods

where t_0 and t_n are the starting and end times of the long-term streamflow record, Q_t is the hourly streamflow at time t and $\bar{Q}$ is the mean streamflow during the long-term record. The philosophy of this modelling framework is to try and introduce covariates that have physical sense in the way in which the catchment will respond to the climate drivers. This covariate is a surrogate for the catchment condition prior to the time when the constituent measurement is taken. An example of Q_m is shown for the Fitzroy catchment in Figure 2 and shows three distinct dry periods (negative values) and four wet periods (positive values). This covariate is also an indicator of whether storages within the catchment are likely to be filling or discharging at the measurement time.

Differentials of Q, q, and Q_m with t are calculated and give the limb of the hydrograph from the sign (positive rising limb and negative falling limb). The magnitude is also a surrogate for the rate at which the area of saturation within the catchment is increasing or decreasing for Q and the groundwater is rising or falling for q. The sign of $\mathrm{d}Q_m/\mathrm{d}t$ ($\mathrm{m}^3\ \mathrm{s}^{-2}$) are an indication of whether the catchment is going into or coming out of a drought or wet period and the magnitude the rate at which this is occurring. The magnitude of $\mathrm{d}Q_m/\mathrm{d}t$ is a measure of how quickly the catchment is changing its condition, while the sign indicates whether the catchment is getting wetter or drier.

Integrals of Q, q, and Q_m with t are calculated for time periods of 1, 2, 5, 10, 20, 50 and 75 days prior to the measurement time. These covariates indicate how much water has come from the catchment leading up to when the measurement is taken and are included so that effects such as the storage and depletion of constituents within the catchment are accounted for. The different length of the integration time is used as different catchments are likely to respond different due to size and catchment characteristics.

Two orthogonal periodic terms, $\sin(2\pi t/86400P)$ and $\cos(2\pi t/86400P)$ were also calculated to determine if an annual periodic signal was present. This term would account for seasonal temperature, biological and possibly management related effects.

3. METHODOLOGY

3.1. Data Preparation

The data consisted of hourly streamflow data, daily baseflow data and measurements of the constituents at points in time at sites in the various catchments. The daily baseflow was converted to hourly baseflow by simple linear interpolation, as the rate of change was slow. Since the streamflow data was assumed constant for the hourly period any constituent measurements that occurred in this hour were averaged and used in the analysis. The constituent data include measurements which were below the detection limit (DT). For these measurements DT/2 was used in the analysis. The data for the DIN was calculated from measured values of the nitrate plus nitrite and the ammonium when both these measurements were made on the sample. The data numbers for the constituents varied from over 1,000 data points in the Tully River catchment at the Euramo site to less than 100 for the Dawson and Comet sub-catchments in the Fitzroy River catchment.

The integrals of Q, q, and Q_m were calculated only when their respective missing data was less than 10% of integral period. The differentials were only calculated if the values were there prior to and after the point of measurement are available.

3.2. Statistical Analysis

A selection process was then developed to select the covariates for inclusion in a multivariate analysis. This consisted of plotting and performing linear regression of the constituent (transformed or not) against a single covariate:

- the constituent (y) versus the covariate (x)
- y versus x^2 or $x^{0.2}$
- $y^{0.2}$ versus x, or x^2 or $x^{0.2}$
- y or $y^{0.2}$ versus $\sin(2\pi t/86400P)$
- y or $y^{0.2}$ and $\cos(2\pi t/86400P)$

From these regressions, any covariate (x) where the regression coefficient (r^2) was greater than or equal to 0.1 was chosen to be included in the multivariate analysis. The power of 0.2 was chosen in preference to a logarithmic transform to avoid the exponential tail term when transforming back to y from $\ln(y)$ when calculating the uncertainty and has been shown to work just as well (McIverney et al., 2017). The chosen covariates and constitutes were then used in the multivariate analysis described below.

The selected covariates were included in an additive general linear model and the method tested all the possible combinations of the selected covariates and the best one was chosen on the basis of the adjusted r^2. The adjusted r^2 was used as it penalises overparameterization to give a model of the following form:

$$z = f(X_1..X_n) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \sum \varepsilon_i \quad (2)$$

where z is constituent (either y or $y^{0.2}$) and X_1 to X_n are the covariates, β_1 to β_n are the coefficients, β_0 is the intercept and ε_i are the error terms which is assumed to be described by a normal distribution. The co-variates can be the untransformed, squared or 0.2 power transformed values. We limited n to five as covariates beyond this number only marginally improved adjusted r^2 and overfits the model, often the number of covariates was less than five.

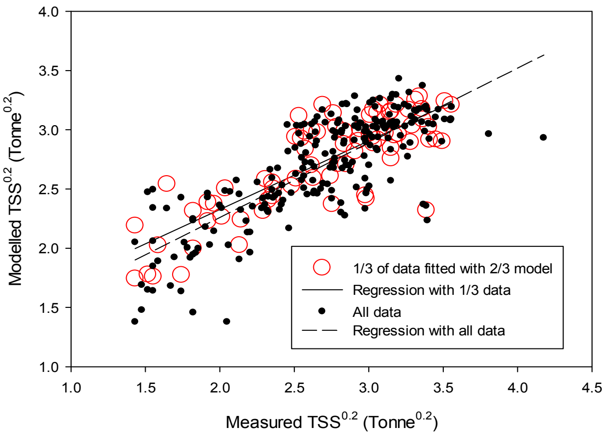

Figure 3. Comparison of statistical model developed from 2/3 of the data and fitted to remaining 1/3 of the data and the statistical modelled developed with all the data and fitted to all the data for the Fitzroy River at Rockhampton.

Table 1. An example of statistical parameters for the multivariate regression models for the constituents for the Tully River site at Euramo (from Cook et al., 2017).

Statistic	TSS	DIN	DON	PN	DIP	DOP	PP
N	1122	1073	1060	1049	1064	1073	1049
Range of values (mg L^{-1})	0.5 - 447	0.0075 - 2.126	0.015 – 0.506	0.015 – 1.39	0.0005 – 0.062	0.005 – 0.23	0.005 – 0.5
r^2	0.389	0.356	0.347	0.454	0.640	0.008	0.253
E	0.673	0.464	0.382	0.271	0.587	-0.796	0.214
dr	0.788	0.585	0.609	0.621	0.745	0.141	0.628
RMSE	33.3	0.159	0.053	0.133	0.004	0.008	0.253
MAE (mg L^{-1})	17.2	0.087	0.040	0.081	0.002	0.007	0.021

3.3. Validation

Various statistical measures of model performance were calculated such as Nash Sutcliffe efficiency (E), regression coefficient, dr (Wilmott et al., 2012), root mean square error (RMSE) and mean absolute error (MAE) (Table 1). The dr statistic is similar to E except it has a lower limit of -1 rather than -∞ and is consider better statistic by Wilmott et al., (2012). MAE needs to be considered with regard to the range of values of the constituent, MAE is 17.2 (mg L^{-1}) for TSS which is 3.8% of the maximum value of 447 (mg L^{-1}), while an MAE of 0.002 (mg L^{-1}) for DIP is 3.2% of the maximum value, for the Tully catchment (Table 1). The multivariate analysis was performed on the whole and a sub-set of 2/3 randomly selected data. The multivariate regression equation from the sub-set was then used to

Table 2. Linear regression parameters for comparison between measured and modelled transformed TSS for the Fitzroy catchment at Rockhampton, for models generated from the full data set, a two thirds subset (Cook et al., 2017) and a further 5 two thirds sub-sets (R1, R2, R3, R4 and R5). Note the original 2/3 model was fitted to the full data set is given here rather than to the 1/3 data set as in the text.

Model	Intercept	Slope	r^2
Full data set	1.17	0.579	0.666
2/3s sub-set	1.44	0.555	0.583
R1	1.09	0.597	0.609
R2	0.95	0.642	0.616
R3	1.08	0.603	0.600
R4	1.03	0.612	0.598
R5	0.97	0.645	0.618

generate constituent values from the covariates from the remaining 1/3 data and compared with the measured values (Figure 3). The linear regression between the measured (*y*) and modelled ($\hat{y}$) transformed TSS are similar for the 1/3 data fit ($\hat{y} = 0.579 + 1.17$, r² = 0.666) and the full data set ($\hat{y} = 0.631y + 1.00$, r² = 0.631). It was only when the 2/3 data set was small in the Dawson River that there was much difference between the statistical models using the 2/3 and full data set. The use of the random selection process may result in selection of samples that are auto-correlated. To verify that, a further five randomly selected sub-sets of 2/3 of the data were selected and models developed from these. These models were used to predict TSS for the Fitzroy catchment at Rockhampton and show only minimal variation in predicted values (Figure 4). Linear regression of the modelled and measured transformed TSS data show that very similar predictions are generated with the further five random sub-sets as well as the original sub-set and with the fully data set (Table 2). It confirms that the randomly selected samples are not necessarily auto-correlated.

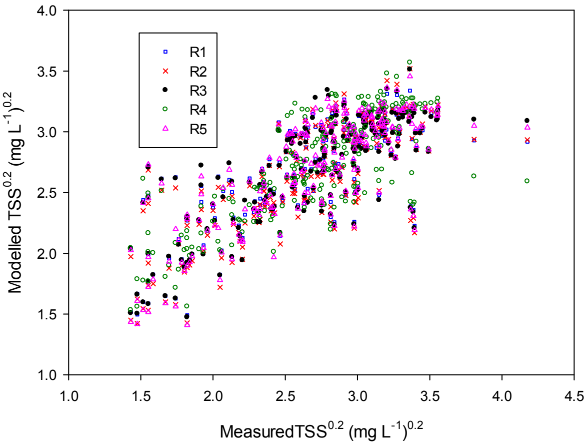

Figure 4. Comparison modelled transformed TSS with measured transformed TSS using models developed from five randomly selected data sub-sets using 2/3of the full data set. For the Fitzroy catchment at Rockhampton.

The annual load data calculated from the hourly concentration data generated from equation (2) multiplied by the hourly streamflow. The load for particulate nitrogen for the Fitzroy catchment at Rockhampton was compared to loads calculated by Department of Science, Information, Technology and Innovation (DSITI) in their various loads reports as well as the estimates by Joo et al. 2012 (Figure 5). The DSITI load is calculated by estimating the area under a plot of the constituent with time. Joo et al. developed a relationship between *Q* and the constituent using statistical methods and the integrating the constituent with time. The load is hindcast back to 2000-2001 using only streamflow data and shows that the model produces loads that are similar to Joo et al. (2012) even though no data prior to 2006 was used in development of our model. The 5th and 95th percentile uncertainty estimates of the concentration were also used to calculate the hindcast loads and show that they bracket the Joo et al. (2012) estimated loads.

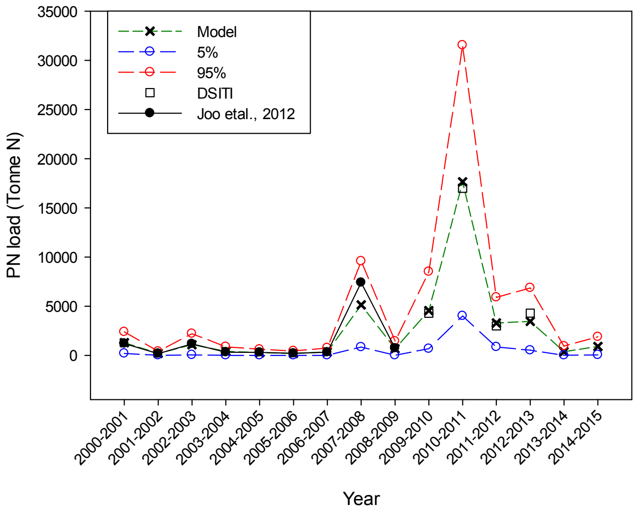

Figure 5. Comparison of annual load of particulate nitrogen (PN), calculated with our statistical model, taken from DSITI loads reports and from Joo et al. (2012). For the Fitzroy catchment at Rockhampton.

4. DISCUSSION AND CONCLUSION

A methodology for the development of a statistical water quality model using an additive general linear model has been developed for seven constituents and based on the conceptual model and the philosophy presented here. Streamflow based parameters are the major drivers of the constituent flow but the use of differentials and integrals of the flow components have proved to be useful, along with the 0.2 power transform. An example is presented in Figure 6 for the comparison of TSS$^{0.2}$ with (dQ/dt)$^{0.2}$. This shows how transform allows the relationship that exists between TSS and dQ/dt to be seen and how if the absolute value of (dQ/dt) is used the rising and falling limbs have a similar relationship with TSS$^{0.2}$. This indicates that it is the magnitude of the change in Q which is more important rather than if the measurements are made on the rising or falling limb of the hydrograph. This has implications for how sampling strategies are designed as it suggests that the sampling on the rising and falling limbs should be at the same density.

The use of integrals of the Q and q is useful in giving an understanding of the recent flow history of the catchment and the likelihood of storage or depletion of the constituent stocks in the catchment. While Q_m gives a longer-term history of the catchment condition in terms of whether the catchment is in a dry or wet period and the sign and value of dQ_m/dt indicates whether the catchment is getting drier or wetter and the rate at which this is happening. We consider that these covariates have added to the ability of this approach to successful estimate the water quality.

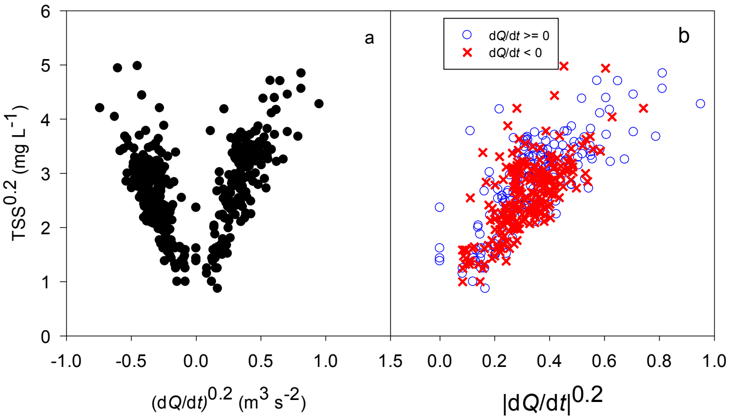

Figure 6. Comparison TSS$^{0.2}$ with a) (dQ/dt)$^{0.2}$ and b) the absolute value |dQ/dt|$^{0.2}$ for the Burdekin catchment at Home Hill.

We have used a general linear model (GLM) rather than a GAMS approach as GAMS models have been shown to over fit the data and we are using the models for forecast. When used for forecasting the splines in a GAMS model would using the end member of the spline which is the least well predicted point (Wood and Augustin, 2002).

This methodology has been shown to work in the catchments of the GBR but given the conceptual model and philosophy that has been used in generating the methodology it should work well in other locations. However, with such a statistical model it is highly dependent on the quality of the data and assumes stationarity in the data. The latter is assumption is likely to fail as the effects of climate change become pronounced and land management changes occur.

The model developed for the Fitzroy catchment at Rockhampton was recently able to give very good forecasts of the load of sediment, nitrogen and phosphorus during cyclone Debbie (Laugesen et al., 2017). This is an extreme event but has shown the robustness and usefulness of the approach. We have been able to hindcast the concentration and loads for the period from 2000 to 2006 with matching of other published loads data (figure 5). No data from this period was used in the development of the models. This gives confidence that the covariates and the approach used here can result in models with good predictive ability.

ACKNOWLEDGMENTS

We thank eReefs for the funding of this project. Freeman Cook would like to thank the Bureau of Meteorology for contract him to participate in this research.

REFERENCES

Abbott, M.B., Bathhurst, J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. (1985). An introduction to the European Hydrologic System (SHE). *Journal of Hydrology* **87**: 45-59.

Arnold, J.G., Srinivasan, R., Muttiah, R.S. and Williams, J.R. (1998). Large area hydrologic modeling and assessment - Part 1: Model development. *Journal of The American Water Resources Association* **34**(1): 73-89.

Cohn TA, Caulder DL, Gilroy EJ, Zynjuk LD and Summers RM (1992). The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake bay. *Water Resources Research* **28(9)**: 2353-2363.

Cohn T.A. (1995). Recent advances in statistical methods for the estimation of sediment and nutrient transport in rivers. *Reviews of Geophysics* **33(S2)**: 1117-1123.

Cook FJ, Jordan PW, Waters DK and Rahman JM (2009). WaterCAST – Whole of catchment model an overview. *In* Anderssen RS, Braddock RD and Newham LTH (Eds) 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Australia 13-17 July, 3492-3499.

Cook, FJ, Khan, U, Laugesen, R, Gnanathikkam, A, Tuteja, N and Bari, M (2017). Water quality forecasting model development using a statistical approach. Freeman Cook and Associates, Bureau of Meteorology 652pp (In Press).

Kroon, F.J., Kuhnert, P,M,, Henderson, B,L,, Wilkinson, S,N,, Kinsey-Henderson, A,E,, Abbott, B., Brodie, J.E. and Turner, R.D.R. (2012). River loads of suspended solids, nitrogen, phosporus and herbicides delivered to the Great Barrier Reef lagoon. *Marine Pollution Bulletin* **65**: 167-181.

Kuhnert, P.M., Henderson, B.L., Lewis, S.E., Bainbridge, Z.T., Wilkinson, S.N. and Brodie. J.E. (2012). Quantifying total suspended sediment export from the Burdekin River catchment using the loads regression estimator tool. *Water Resources Research* **48**: W04533.

Laugesen RM, Tuteja NK, Amirthanathan GA, Kent D, Hasan M, Khan U, Bari M (2017). Forecasts and simulations of river water quantity and quality caused by Tropical Cyclone Debbie over Fitzroy River Basin, 22nd World IMACS Congress and MODSIM17 International Congress on Modelling and Simulation, Hobart, Australia 3-8 December

McInerney D, Thyer M, Kavetski D, Kuczera G and Lerat J (2017). Evaluation of approaches for modelling heteroscedasticity in the residual errors of hydrological predictions *Water Resources Research*, 53:2199-2239.

Miller, C.R. (1951). Analysis of flow-duration, sediment-rating curve method computing sediment yield. Report, U.S. Bureau of Reclamation, Denver, Color., 15pp.

Robson, B.J. and Dourdet, V. (2015). Prediction of sediment, particulate nutrient and dissolved nutrient concentrations in a dry tropical river to provide input to a mechanistic coastal water quality model. *Environmental Modelling & Software* **63**: 97-108.

Rustomji, P. and Wilkinson, S.N. (2008). Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves. *Water Resources Research* **44(9)**: W09435.

Turner, R., Huggins, R., Wallace, R., Smith R, Vardy, S. and Warne, M.St.J (2012). Sediment, Nutrient and Pesticide Loads: Great Barrier Reef Catchment Loads Monitoring 2009-2010. Department of Science, Information Technology, Innovation and the Arts, Brisbane.

Wilmott, C.J., Robeson, S.M. and Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology* **32**: 2088-2094.

Wang, Y-G., Kuhnert, P. and Henderson, B. (2011). Load estimation with uncertainties from opportunistic sampling data – A semiparametric approach. *Journal of Hydrology* **396**: 148-157.

Wood, S.N. and Augustin, N.H. (2002). GAMS with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling* **157**: 157-177.