BMT Eastern Australia Pty Ltd
Level 8, 200 Creek Street
Brisbane Qld 4000
Australia
PO Box 203, Spring Hill 4004

Tel: +61 7 3831 6744
Fax: + 61 7 3832 3627

ABN 54 010 830 421

www.bmt.org

1 Background

BMT have previously submitted an expression of interest (EoI) as part of the Department of Environment and Science's RFQ titled Queensland Water Modelling Network RDI Priorities. This document constitutes a detailed proposal by BMT, Healthy Land and Water and University of Western Australia.

The state of our catchments plays an important role in determining the health of some of our most sensitive coastal ecosystems. Catchment models like eWater's SOURCE are often used to model catchment runoff from applied precipitation alongside concentrations of water quality constituents. Predictions of loads generated from these catchments are often used to inform management practices and critical investments to maintain the health of downstream ecosystems. Predictions of nutrients from SOURCE have the following major drawbacks:

- The predictions are based on simplistic load generation models within SOURCE that do not adequately capture the complex, non-linear relationship with the actual physical environment. In the most common setup, constant values are specified for Event Mean Concentrations (EMCs) and Dry Weather Concentrations (DWCs), and the model switches between the two values depending on the case.
- In its current form SOURCE has only been used to provide predictions for Total Nitrogen and Total Phosphorus. An understanding of water quality constituents like particulate, labile, inorganic and organic nitrogen and phosphorus is useful in improving decision making around catchment management and quantifying impacts on downstream ecosystems.
- While parameter uncertainty associated with SOURCE and its setup has been extensively researched, an understanding of the overall model uncertainty with respect to observations of parameters with non-linear behaviour is lacking. This is especially critical when outputs from SOURCE are used to force other models.

Data driven models and ensemble machine learning techniques can potentially be used to improve predictions of water quality constituents. This project aims at building up on existing knowledge with respect to application of machine learning methods to predict nutrient concentrations from catchments and quantify uncertainty associated with SOURCE predictions using the same methods. The following sections describe our methodology.

2 Project Methodology

2.1 Overview of the models

The proposed study envisages using a variety of approximation models and techniques to predict nutrient generation from catchments. Results from different models will be compared amongst each other to arrive at the best method.

2.1.1 Linear and Multivariate Weighted Regression models

Linear models are the most commonly used tool to describe concentration-discharge (C-Q) relationships (Hirsch, Moyer and Archfield, 2010). Often log transformations are used and the final linear model can be described as:

$$\log(C) = \beta_0 + \beta_1 \log(Q)$$

The fitted slope β_0 represents the base concentration in the stream, while β_1 describes the relationship between hydrological and biogeochemical data. The simple linear model can be further extended into a multivariate weighted regression model called Weighted Regressions on Time, Discharge and Season (WRTDS) model:

$$\log(C) = \beta_0 + \beta_1 \log(Q) + \beta_2 JD + \beta_3 \sin(JD) + \beta_4 \cos(JD) + \varepsilon$$

JD represents the Julian Day and ε represents the unexplained variability in the data. $\beta_2 JD$ represents the long-term, annual trends, while $\beta_3 \sin(JD)$ and $\beta_4 \cos(JD)$ describe seasonal variations in stream concentrations.

The WRTDS model provides significant improvement over the linear model by accounting for seasonal and long term trends within the data. Unlike the linear model whose parameters are fixed in time, the parameters in WRTDS get adjusted gradually throughout the Q, JD space. This is done using weighted regression for the estimation of log(C), where the weights on each observation are based on the following distances between the observation (Q_o, JD_o) and the estimation point (Q_i,JD_i):

- Time distance between JD_o and JD_i
- Seasonal distance between the time of year at JD_o and the time of year at JD_i.
- Discharge distance between $\log(Q_o)$ and $\log(Q_i)$

2.1.2 Random Forest and Gradient Boosting Machines

Random Forest (RF) is an ensemble machine learning technique that can be used for both regression and classification tasks. RF uses multiple decision trees whose predictions are aggregated using bootstrap aggregation, also commonly called bagging. This involves using a different data sample for training each tree with replacement from the original set.

Gradient Boosting Machines (GBM) is a framework where decision trees or any form of base learners are recursively improved in terms of their prediction quality. The first step involves fitting the model to the data, then fitting a model to the residuals and getting the new model by combining the previous model and the residual model. The iteration process is described as:

$$F(x) = F_1(x) \rightarrow F_2(x) = F_1(x) + h_1(x) \ldots \rightarrow F_M(x) = F_{M-1}(x) + h_{M-1}(x)$$

$h_m(x)$ simply represents a model of the residuals while $F_m(x)$ is a model of the predictions from the base learner.

The underlying base learner used typically with RF and GBM methods is the Classification and Regression Tree (CART). As the name suggests they can be used for both classification and regression tasks. Assuming R^d is the data space, the data can be split into K disjoint subspaces $\{R_1, R_2 \ldots, R_k\}$ where each $R_j \subset R^d$. The same decision/prediction is made for all $x \in R_j$ for each feature subspace. The following algorithm is used for generating regression trees (Breiman, 2017):

- Begin with the first feature subspace.
- For each feature j = 1,…,d and for each value v $\in$ R on which a split is possible:
    - Split the dataset:

$$I_L = \{\text{X1}, x_i^a \ < \ \text{v}\} \text{and } I_R = \{\text{X2}, x_j^a \ \geq \ \text{v}\}$$

    - Estimate the average y for each node using $\overline{y_R}\ and\ \overline{y_L}$.

$$\overline{y_L} = \frac{\sum_{i \in I_L} y_i}{|I_L|}\ and\ \overline{y_R} = \frac{\sum_{i \in I_R} y_i}{|I_R|}$$

    - Estimate the quality of the split by calculating the squared loss.

$$squared\ loss = \sum_{l=0}^{L} (y_l - \overline{y_L})^2 + \sum_{l=0}^{L} (y_r - \overline{y_R})^2$$

- Choose the split with minimal loss.
- Do the process recursively on both children.

Both RF and GBM are essentially ensemble methods and combine a number of regression trees to make predictions. The averaged values in the terminal node of each tree are treated as the prediction of that tree and the average of all the trees is taken as the final prediction.

2.2 Generation of Flow Data

The first step will involve the separation of the Baseflow and Quickflow components at each of the monitoring locations. The three passes filtered method will be applied for baseflow separation. The quickflow will be calculated as below:

$$QF_i = \alpha QF_{i-1} + (Q_i - Q_{i-1}) \frac{1 + \alpha}{2}$$

QF_i is the filtered quickflow for the ith sampling instant and QF_{i-1} is the filtered quickflow for the previous sampling instant. α will be assumed to be 0.925 . The baseflow will be calculated as BF = Q - QF.

The baseflow separation will be done using the EcoHydRology package in R. The generated baseflow, quickflow, total flow and rainfall will be transformed into lagged data (averaged values over the previous days) to capture any short-term impacts of different water pathways and rainfall on stream nutrients. JD, cos(JD) and sin(JD) will also be calculated for RF and GBM to include seasonal and long-term impacts.

2.3 Hybrid Models

Hybrid models are a combination of multiple prediction sources combined together in a broader machine learning framework to improve overall predictions. In the context of this project three different forms of hybridisation will be tested.

The first involves the pre-generation of predictions from a standalone machine learning model (GBM or RF) and then these are combined with the observations, lagged hydrological data, lagged rainfall data and temporal data to make the final predictions (Wang, Hipsey and Oldham, 2019). This method has already been tested on catchments in Western Australia by Prof. Carolyn Oldham, A/Prof Matt Hipsey and Benya Wang as part of their doctoral research.

The second method involves replacing machine learning predicted data in the first step with predictions from SOURCE. The predictions from SOURCE, observations, lagged hydrological data, lagged rainfall data and temporal data are combined in a RF and GBM method. This method has been used previously by Isik *et al.*, 2013 to improve daily flow predictions using Soil and Water Assessment Tool (SWAT) and Artificial Neural Networks.

The third method involves the use machine learning models to predict the error in SOURCE predictions using observation data and other lagged hydrological data, lagged rainfall data and temporal data. This method can be used to model the bias in SOURCE predictions and helps quantify uncertainty in SOURCE predictions. This method has been used to improve groundwater predictions from MODFLOW (Xu *et al.*, 2012).

2.4 Modelling Process

The observed data points will be divided into the training dataset and the testing dataset. Different models will be built and tuned on the training dataset and the testing dataset will be used for the final test. Cross-validation (CV) will be done on the training dataset to tune the model parameters.

The WRTDS model will be run using the Exploration and Graphics for River Trends (EGRET) package in R to produce daily concentrations of six nutrient species - TN, TP, DON, DOP, DIN and FRP. Default settings specified in the user guide will be used.

The RF and GBM models will be built using the H_2O package in R.

The hybrid models will be setup using the methods outlined in section 2.3. The performance of all the models - LM, WRTDS, RF, GBM, RF-RF hybrid, GBM-GBM hybrid, SOURCE-RF hybrid, SOURCE-GBM hybrid, SOURCE-RF error prediction and SOURCE-GBM error prediction will be checked against the test data. To assess the prediction uncertainty of the models the process will be repeated multiple times with different training and test datasets each time. WRTDS will not be repeated as part of this process and Leave-one-out cross-validation (LOOCV) will be done with its outputs.

2.5 Evaluation Metrics

In this study, the root mean squared error (RMSE) and the Nash-Sutcliffe model efficiency (MEF) will be used to compare model performance. The RMSE is a measure of overall error between the predicted and measured data and returns an error value with the same units as the data, which is given by the following equation:

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

n is the number of observations. RMSE varies from 0 to +∞, and a perfect model would have RMSE of 0. The MEF is a dimensionless 'goodness of fit' measure which can vary from -∞ to 1, with a value of 1 indicating perfect fit and 0 indicating that the mean of the measured value performs as well as the model. The MEF is calculated as:

$$MEF = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

$\bar{y}_i$ is the mean of the observed values.

3 Data Sources

The project is primarily based on water quality observation data collected by Healthy Land and Water. Five catchment water quality monitoring sites near existing Department of Natural Resources and Mines (DNRM) gauges have been set up and data has been collected between 2014 and 2019. Figure 3-1 shows the locations of these monitoring sites and their positions with respect to the major catchments. TN, PON, NO_x, NH_3, TP, DOP, POP and FRP have been collected as part of the monitoring.

Hourly stream gauge data will be sourced from existing DNRM gauges and hourly rainfall data will be sourced from Bureau of Meteorology (BoM) gauges.

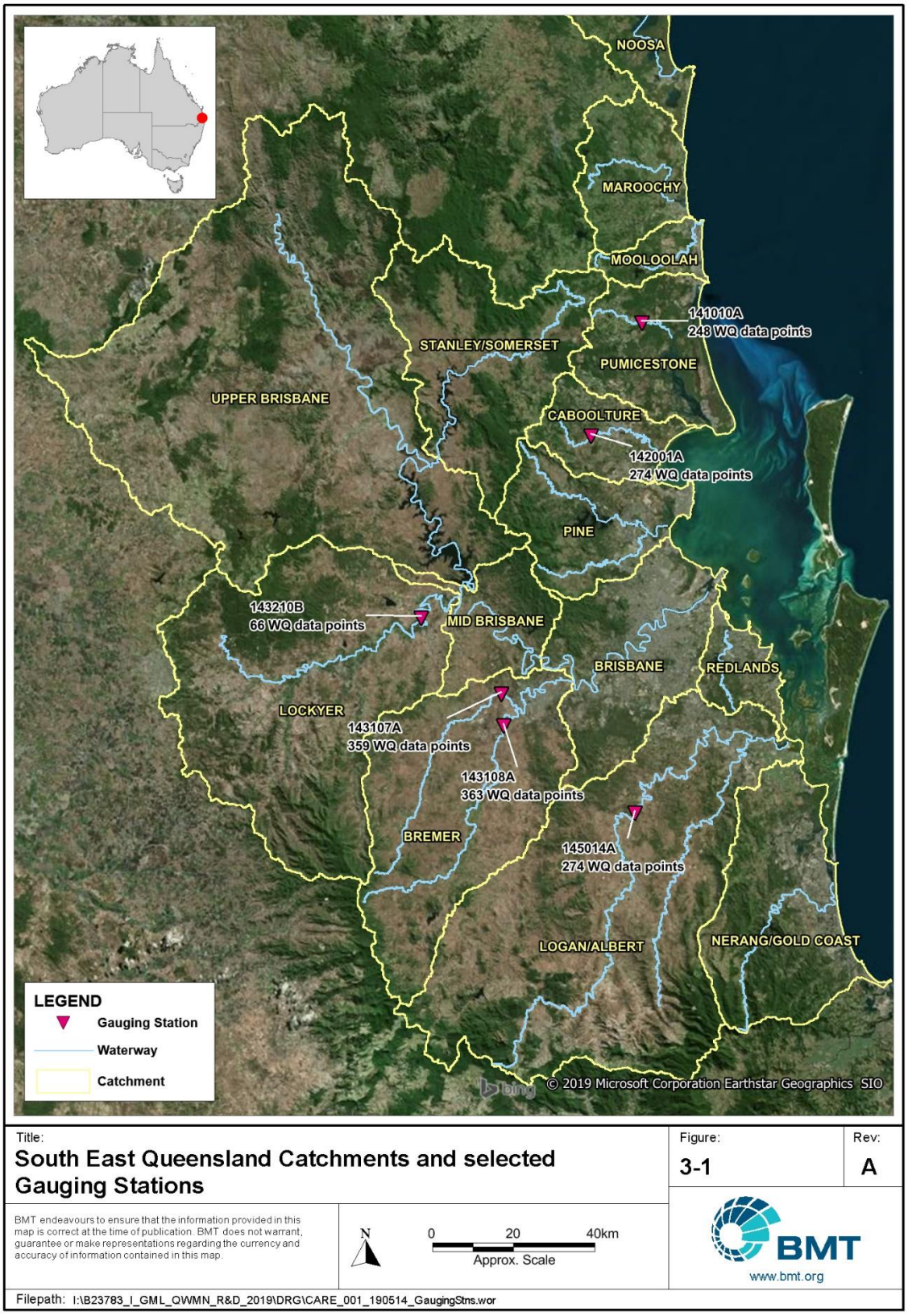

Figure 3-1 South-east Queensland catchments and selected gauging stations

4 Practical Application Pathway

Nutrient budgets and estimation of catchment loads is an area of interest for stakeholders like Healthy Land and Water. The data outlined in the previous section is collected as part of a dedicated program called the South-East Queensland Catchment Loads Monitoring Program (SEQ CLMP). The observed concentration data is used with interpolated gauge flow data to calculate the load generated by each catchment.

The data collected is often sparse and interpolations between observations introduce significant uncertainty in calculated predictions. Using SOURCE could be one possible way of avoiding interpolations, however as mentioned before the intrinsic water quality model setup does not capture the complex, non-linear interactions with the environment.

The proposed data driven model as part of this project can be used to improve load predictions from catchments.

Another application is the linked SOURCE - TUFLOW FV models that are used to predict water quality in SEQ estuaries. Currently, there is no quantification of uncertainty involved when linking the two models. Method 3 outlined above will specifically address this point.

The proposed solutions can also be applied elsewhere, especially the Great Barrier Reef catchments where nutrient load management and their prediction is an important issue.