26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Modelling upstream water quality from downstream observations in the Muttama catchment

Vincent S Khang[1], Clara Grazian[2], R. Willem Vervoort[1]

School of Life and Environmental Sciences, The University of Sydney, Australia[1], School of Mathematics and Statistics, The University of Sydney, Australia[2]
Email: ukha0460@uni.sydney.edu.au

Abstract: Monitoring the health of river catchments through water quality sampling is important for environmental protection. However, resources are generally limited, resulting in fewer monitoring samples in regional river systems. This consequently leads to sparse spatial and temporal water quality datasets, creating an incomplete view of the catchment health and this hinders effective management of the catchment. To address these concerns, space-time models and interpolation techniques can be employed to predict water quality under data-sparse conditions, for example to manage and minimise potential environmental risks within a catchment.

In this study, space-time modelling approaches for water quality will be investigated in the context of the Muttama catchment in New South Wales, Australia. This paper will demonstrate: (i) how a load prediction model can address the data sparsity: and (ii) how generalised additive models (GAMs) can predict concentrations of water quality parameters at upstream sites when trained on the most downstream site (Coolac).

The Muttama catchment dataset was obtained from Vervoort et al. (2025). Interpolation of the concentration within this dataset used a load prediction model which included seasonality. Using a logarithmic transformation, the load equation increased the number of complete datasets for training and testing from 124 to 227. The chemical signatures used for this interpolation were $\{\mathrm{Cl^-}, \mathrm{Ca^{2+}}, \mathrm{Mg^{2+}}, \mathrm{Na^+}, \mathrm{SO_4^{2-}}, \mathrm{HCO_3^-}, \mathrm{K^+}\}$ and these had a $R^2 = \{0.615, 0.263, 0.261, 0.218, 0.189, 0.171, 0.168\}$ respectively. The load prediction model was statistically significant with $0.001 < \mathrm{p} \leq 0.007$ across the different concentrations. The chemical covariates used for the model were chosen from the correlation-based feature selection (CFS) after performing a correlation analysis on electrical conductivity (EC) and the chemical concentrations. It was found that the optimal subset of non-correlated chemical covariates were $\{\mathrm{Cl^-}, \mathrm{HCO_3^-}, \mathrm{SO_4^{2-}}\}$, with a CSF score $S_{CFS} \approx 0.874$ and these chemical concentrations were used to train the model.

Separation of the dataset via k-mean clustering also improved the model's performance by allowing for weight specific conditioning to occur for clusters that exhibit higher EC values on average. Each cluster was assigned a different set of weights on the covariates, and this was integrated into a gamma distributed generalised additive model (GammaGAM). The GammaGAM was used to predict EC upstream and the model performed well for low $EC \leq 2500\ \mu S.cm^{-1}$, but struggled with $EC > 2500\ \mu S.cm^{-1}$. The GammaGAM performance greatly improved when only non-correlated covariates were used in the model and site-specific proxy covariates (e.g. distance and elevation from Coolac) were removed.

The results show that the load prediction model equation can be a powerful tool for interpolation, especially within the context of data sparsity. These findings highlight the application of GAM when used on regional river catchments where observed data is sparse and unevenly distributed.

REFERENCES

Vervoort, R. W., van Ogtrop, F., Tambrchi, M., Akter, F., Buzacott, A., Lessels, J.,Bishop, T. (2025). A detailed streamflow and groundwater salinity dataset for Muttama Creek Catchment, NSW, Australia. Earth Syst. Sci. Data Discuss., 2025, 1-26. https://doi.org/10.5194/essd-2025-105

Keywords: *Hydrology, geostatistics, surface water.*

