



OPEN

# Explainable AI approach with original vegetation data classifies spatio-temporal nitrogen in flows from ungauged catchments to the Great Barrier Reef

Cherie M. O'Sullivan<sup>1</sup>✉, Ravinesh C. Deo<sup>2,3</sup> & Afshin Ghahramani<sup>1,4</sup>

Transfer of processed data and parameters to ungauged catchments from the most similar gauged counterpart is a common technique in water quality modelling. But catchment similarities for Dissolved Inorganic Nitrogen (DIN) are ill posed, which affects the predictive capability of models reliant on such methods for simulating DIN. Spatial data proxies to classify catchments for most similar DIN responses are a demonstrated solution, yet their applicability to ungauged catchments is unexplored. We adopted a neural network pattern recognition model (ANN-PR) and explainable artificial intelligence approach (SHAP-XAI) to match all ungauged catchments that flow to the Great Barrier Reef to gauged ones based on proxy spatial data. Catchment match suitability was verified using a neural network water quality (ANN-WQ) simulator trained on gauged catchment datasets, tested by simulating DIN for matched catchments in unsupervised learning scenarios. We show that discriminating training data to DIN regime benefits ANN-WQ simulation performance in unsupervised scenarios ( $p < 0.05$ ). This phenomenon demonstrates that proxy spatial data is a useful tool to classify catchments with similar DIN regimes. Catchments lacking similarity with gauged ones are identified as priority monitoring areas to gain observed data for all DIN regimes in catchments that flow to the Great Barrier Reef, Australia.

Communicating catchment influences towards the ecology of the receiving environment is enhanced by water quality simulation tools. Customising water quality simulation models to the catchment they represent is essential for limiting uncertainty in results and maintaining trust in land use decisions they aim to inform<sup>1,2</sup>. Model design and development, referred in here as customisation, is achieved by using observed water quality data from gauging stations for design and verification of models<sup>3</sup>. However, many water catchments globally are ungauged, and a lesser proportion of those have corresponding water quality data to inform model customisation. Techniques to overcome such data voids in ungauged areas are necessary<sup>4,5</sup>. Methods to simulate flows in ungauged areas are well researched<sup>6,7</sup>, however, refinement of methods that simulate nutrients in ungauged areas remained unresolved. This knowledge gap in water quality modelling needs addressing to best inform anthropogenic nitrogen management, and to demonstrate progress to the 2030 UN Nations Sustainable Development Goals commitment to reduce land-based nutrients that enter the oceans<sup>8</sup>. This has relevance for the Great Barrier Reef World Heritage Area where over ~20% of the terrestrial drainage area is ungauged, and nutrient balances are critical for the reef's health<sup>9,10</sup>. Logical explainability in nutrient models for ungauged areas can support communications and enable more responsive water quality improvement investments<sup>11,12</sup>.

<sup>1</sup>University of Southern Queensland, Toowoomba, QLD 4350, Australia. <sup>2</sup>School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD 4300, Australia. <sup>3</sup>Center for Applied Climate Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia. <sup>4</sup>Present address: Department of Environment and Science, Queensland Government, Rockhampton, QLD 4700, Australia. ✉email: Cherie.O'Sullivan@unisq.edu.au

For both data-driven and process-based models that simulate water quality, observed water quality and quantity data, as well as a comprehensive understanding of catchment characteristics are required<sup>13</sup>. Data driven water quality models are useful to forecast water quality output, but water flows and water quality must be known a priori to develop covariates<sup>14,15</sup>. In contrast, process-based models use physical and empirical principles and can be established for catchments lacking observed water quality data. In ungauged areas, data is donated to ungauged catchments from the most similar gauged ones<sup>16,17</sup>. Alongside these traditional water quality modelling approaches, deep learning, particularly in the revised forms of Artificial Neural Networks, has been relatively successful in simulating water quality, including nitrogen, without the need for prior established principles<sup>18,19</sup>. As a subsector of Deep Learning, Artificial Neural Networks have demonstrated the ability to recognise patterns in input datasets, classify them, and establish algorithms to match target data. The merits of ANN are demonstrated to forecast and extend non-linear water quality data within respective catchment datasets<sup>20</sup>, but their application to inform scenario simulation, and hence land management decisions, which is the benefit of process-based models, is lacking<sup>21</sup>.

To exploit the benefits and overcome the drawbacks of each data driven vs process based model approach, the coupling of machine learning models such as ANN with process-based approaches can be performed to provide benefits of transfer learning<sup>22,23</sup>. However, machine learning models that incorporate process considerations for water quality modelling are disproportionately underrepresented in many research articles<sup>22,23</sup>. Additionally, where applied to ungauged areas, low landscape heterogeneity between drivers for the constituent being simulated is necessary<sup>21,24</sup>. While variations in patterns of nutrients are observed across gauged catchments that drain to the Great Barrier Reef<sup>15,25,26</sup>, methods for classifying those catchments to the most similar ungauged catchments that drain to the Great Barrier Reef, based on similarity of nitrogen drivers, are unexplored.

In terrestrial landscapes, Dissolved Inorganic Nitrogen molecules are influenced by decomposers, vegetation uptake, nitrogen fixing bacteria etc., which change depending on a unique combination of physical and biological influences at each location<sup>27,28</sup>. The fluxing nature of these biotic processes mean catchment similarities for drivers of DIN differ from the abiotic drivers of flow have therefore been complicated to quantify<sup>22,29</sup>. Variability in drivers of DIN affect the consistency of water quality modelling of ungauged areas<sup>25</sup>. This disparity between biotic and abiotic influence on nitrogen drivers means classical classification approaches that only use physical similarities miss the influence of all biologically influenced differences that may exist between catchments. Spatio-temporal variability in nutrient drivers can be represented in catchment models by the natural physical drivers of geology, aspect, topography, climate etc., as well as land use to represent anthropogenic impacts, including standard fertiliser application rates which affect DIN<sup>25</sup>. Our earlier studies found Original Vegetation is a proxy dataset for the residual biotic responses to these, and any other unaccounted-for drivers that can parsimoniously classify catchments for DIN, and identify the classification drivers using explainable artificial intelligence, (XAI)<sup>30,31</sup>.

Explainable artificial intelligence, (XAI) has outstanding capabilities to highlight the influential variables in machine learning algorithms, however, performance criteria for the corresponding ANN models are likely to vary unpredictably with changes to model architecture and scenarios<sup>32</sup>. Established process-based models instead can be customised to respective catchments using regionalised parameter data, enabling trials of different land management scenarios<sup>18</sup>. This technique has been effective for water quality constituents driven by abiotic processes, which result in consistent performance<sup>6</sup>, and so is pragmatic for the purpose of informing land management decisions. Despite the acceptable track record for process-based models, the suitability of parameter transfer for constituents with biotic process drivers is still lacking, and studies regarding the spatio-temporal scales are necessary<sup>33–35</sup>. We found earlier that original vegetation can be a proxy for matching gauged catchments with dynamic DIN patterns<sup>26</sup>. However, no approach has yet been developed that matches ungauged to gauged catchments for DIN similarities, which would be beneficial for models that transfer data across catchments with similar processes.

This study extends our previous XAI-SHAP<sup>30,36</sup> approaches to match the currently ungauged to gauged catchments that flow to the Great Barrier Reef using mapped spatial data as a proxy for DIN. Mapped spatial data is useful because it provides data for all areas of the Great Barrier Reef catchments where water quality data is lacking<sup>26</sup>. In this study we verify the classification results by building and applying an ANN-WQ simulator to compare changes in simulation performance criteria for a case study catchment, under various dataset arrangement scenarios. Our earlier studies found dominant original vegetation data features may provide guidance to the part of the hydrograph that is relevant to consider for matched catchments and that it is a useful proxy to group gauged catchments that flow to the Great Barrier Reef to three DIN response categories<sup>30</sup>. In this study we evaluate whether our previous method is extendable to all ungauged catchments that flow to the Great Barrier Reef, and undertake a case study to verify its suitability as a proxy for DIN classification. Our verification case study aims to confirm catchments classified together based on original landscape variables also have transferrable water quality responses that can be exploited to simulate DIN.

The hypotheses investigated here are: (1) Original vegetation spatial features found to be a proxy for DIN discharge from gauged catchments in our previous studies<sup>26,30</sup> can be used to match gauged catchments to ungauged catchments that also flow to the Great Barrier Reef. (2) An ANN-WQ simulator trained using predictor variables of original vegetation, coupled with flow data characterised to match with DIN targets will achieve superior performance compared to an ANN-WQ simulator trained to simulate DIN using non-categorised flow data only.

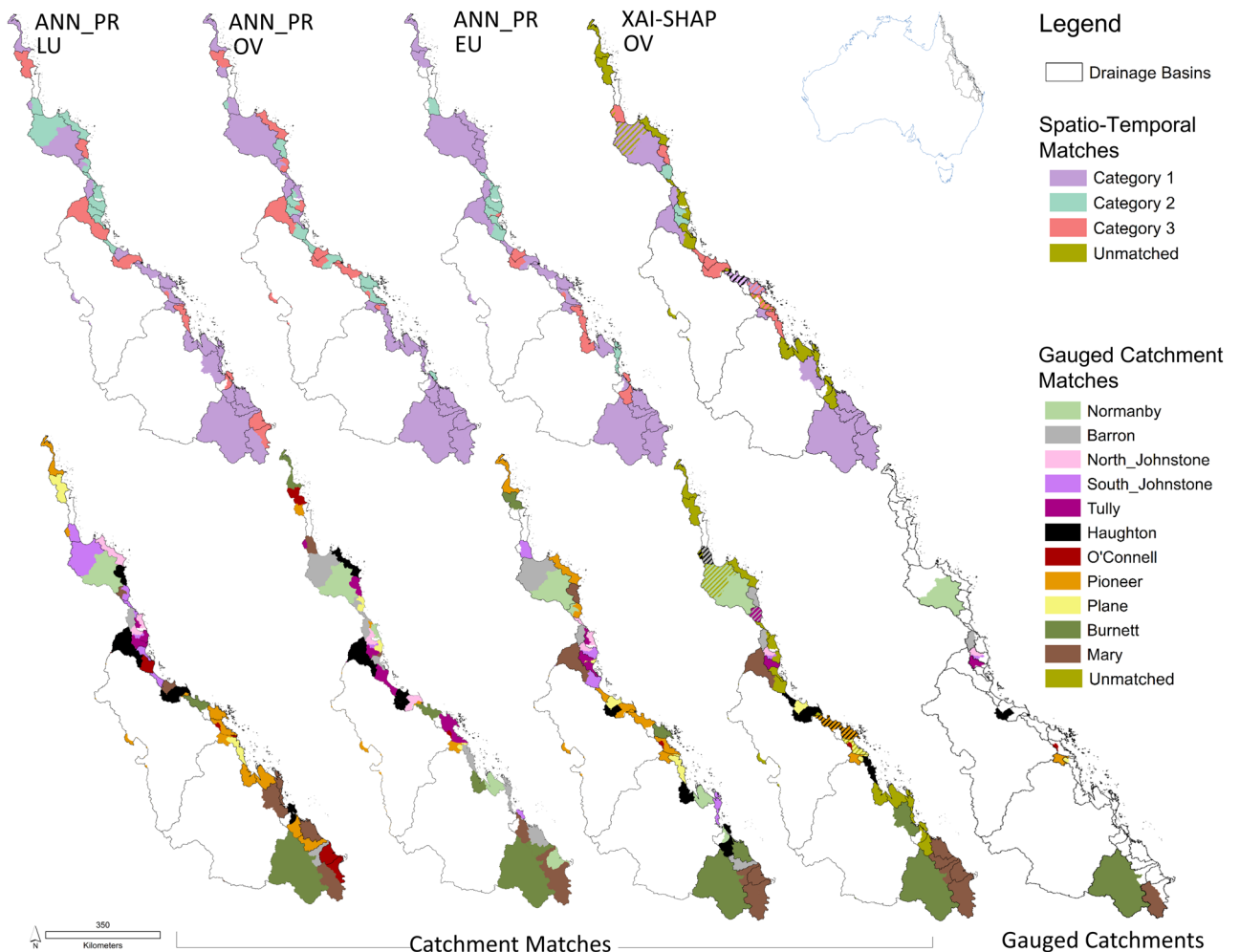
(3) The trained ANN-WQ simulator can simulate DIN in an unsupervised scenario for a pseudo-ungauged case study catchment matched based on the spatial proxy data and achieve satisfactory performance criteria to verify the suitability of the catchment match approach. For this study, pseudo-ungauged means a gauged catchment, with the same data collection method as the other gauged catchments, but intentionally omitted from previous research that informed this study. DIN data for the pseudo ungauged catchment is used here for hypothesis validation purposes only.

The present study therefore aims to create an XAI approach for considering original vegetation data classification as the proxy for spatio-temporal nitrogen patterns in ungauged catchment flows for the specific case of the Great Barrier Reef in Australia.

## Results

### ANN-PR matches

Apart from the Mary Catchment, the results show that ungauged portions of gauged catchments do not necessarily classify together, and catchments do not necessarily classify with their nearest neighbours (Fig. 1). Catchment matches varied for each spatial dataset evaluated, and translation of those results to classify catchments based on corresponding DIN response Categories also varied (Table 1). While Category 2 matched catchments generally clustered together spatially, Category 3 matched catchments contrasted with distributions only north of Plane for the Original Vegetation (OV) dataset compared to further south where Land use (LU) variability was included independently or embedded within the Ecounit (EU) data. This indicates that the catchments in the different datasets show different spatial characteristics. For example, the catchments that matched Category 2 tended to be clustered together, while the catchments that matched Category 3 showed more variability when the LU dataset, which represents anthropogenic, in contrast to natural biotic response to environmental influences, was included.



**Figure 1.** Catchment matches using ANN-PR and XAI-SHAP approach to identify ungauged catchment similarities to gauged catchments using spatial dataset (OV original vegetation, EU ecounit, LU land use). Top row shows the spatio-temporal category of the matched gauged catchment based on the gauged catchment allocation derived from our previous works<sup>30</sup>, bottom row shows the matched catchment. Colours represent the gauged catchment as listed in the legend. Results show high variation between each dataset. Maps created by author using ArcMap 10.8.1, gauged catchments<sup>10</sup> supplied, Drainage Basins<sup>37</sup> licenced under a Creative Commons—Attribution 3.0 Australia licence (CC BY 3.0 AU). © State of Queensland (Department of Environment and Science) 2023.

Category	Scenario	Performance						Best hrs					
		MSE	R2	NSE	d	RMSE	pde	MSE	R2	NSE	d	RMSE	pde
All	Control	0.004	0.390	0.130	0.336	0.060	<b>-9.323</b>	435	879	435	113	435	14
	EULUOV	0.002	0.845	0.634	0.809	0.048	32.780	920	976	920	989	920	2
	LU	0.005	0.361	0.116	0.279	0.073	81.593	893	840	893	926	893	14
	EU	0.006	0.324	0.093	0.219	0.074	12.930	607	848	607	449	607	3
	OV	0.005	0.340	0.111	0.268	0.074	6.920	911	911	911	485	911	27
Category 1	Control	0.008	<b>0.686</b>	4.988	0.454	0.089	<b>-5.208</b>	653	170	653	170	653	363
	EULUOV	0.001	0.613	0.252	0.532	0.036	<b>-10.058</b>	621	988	621	948	621	3
	LU	0.001	0.512	0.247	0.568	0.036	<b>-0.703</b>	260	439	260	439	260	28
	EU	0.001	0.581	0.252	0.570	0.036	14.235	587	947	587	332	587	10
	OV	<b>0.001</b>	0.649	<b>0.375</b>	<b>0.637</b>	0.033	-1.830	975	867	975	827	975	16
Category 2	Control	<b>0.002</b>	<b>0.857</b>	<b>0.688</b>	<b>0.846</b>	<b>0.045</b>	12.451	990	990	990	990	990	8
	EULUOV	0.003	0.823	0.609	0.799	0.050	42.345	949	795	949	949	949	506
	LU	0.002	0.837	0.637	0.814	0.048	34.739	903	619	903	951	903	3
	EU	0.002	0.828	0.619	0.794	0.049	<b>12.045</b>	902	941	902	973	902	13
	OV	0.002	0.845	0.634	0.809	0.048	32.780	920	976	920	989	920	2
Category 3	Control	0.002	0.971	0.918	0.947	0.041	-0.250	426	946	426	426	426	380
	EULUOV	0.001	0.974	0.929	0.956	0.038	0.012	990	990	990	558	990	370
	LU	0.002	0.973	0.911	0.948	0.043	<b>-0.005</b>	908	614	908	908	908	457
	EU	<b>0.001</b>	<b>0.976</b>	<b>0.941</b>	0.955	<b>0.035</b>	0.053	269	269	269	269	269	452
	OV	0.001	0.975	0.938	<b>0.958</b>	0.036	-0.009	905	408	905	905	905	502

**Table 1.** Performance evaluation of ANN-WQ simulator for the Gauged Catchment scenarios. Shading intensity represents performance over all scenarios tested for each criteria. Scores styled **bold** are the best performing metric for each category, scores styled *italic* fail to meet minimum satisfactory performance criteria. Training datasets discriminated by category influence DIN simulation results.

### Variable feature independence

In this study, matching variable deviations using the XAI-SHAP approach method<sup>30</sup> revealed that every catchment had a unique combination and weighting of deviated features. The same combinations of top XAI-SHAP 10% floristic structure variables did however match the most similar gauged catchment and group them to Categories based on the combination of deviated variables. It also revealed catchments that did not share the same combinations of deviated variables. Grouping the deviated variables by landform and vegetation descriptors in the Original Vegetation dataset allowed for 20 of the 37 pseudo/ungauged catchments to be matched to individual gauged ones, while 9 catchments were not matched to another ungauged or gauged catchment or spatio-temporal category. Of those, unable to match to gauged catchments XAI-SHAP results facilitated four closely matched groups to be identified.

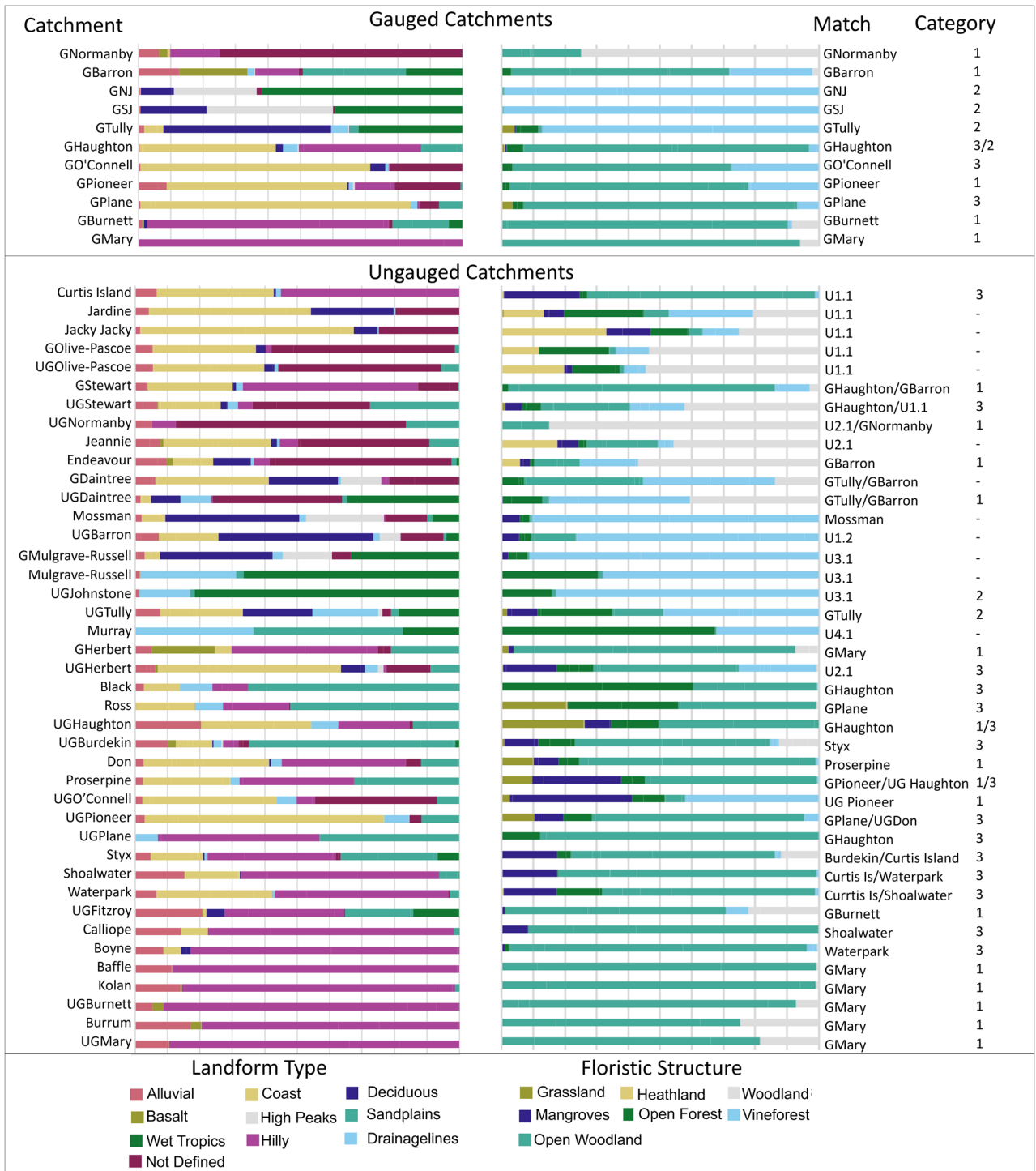
Variable combinations only occurring in ungauged catchments and not in the gauged ones include: hilly alluvial with basalt, health land with sandplains and coast, or mangrove landform structures, as well as additional combinations of vineforest with woodland drainage, or open forests combined with grassland and open woodlands (Fig. 2).

### ANN-WQ simulator performance

The most notable observation was that the combination of catchments included in the training datasets influenced the unsupervised performance of the ANN-WQ simulator (Table 1). When the ANN-WQ simulator was trained using data for individual catchments, simulations were only able to be generated in the unsupervised environment for the Wet Tropic catchments of Tully, and North and South Johnstone. Flatline simulations were observed in the unsupervised simulator environment for all other catchments, despite their adequate training performance (Supplementary Material Fig. SF1).

In contrast, training using data grouped from multiple catchments generated non flatline results for all scenarios. Satisfactory to very good performance for all metrics were achieved for all spatial dataset combinations grouped and discriminated in spatio-temporal Category 2 and 3. Except for their unsatisfactory Nash Sutcliffe Efficiency (NSE) performance, datasets grouped and discriminated to spatio-temporal Category 1 also achieved

### Top 10% XAI-SHAP deviations



**Figure 2.** Top 10% XAI-SHAP deviations for landform and flora sub-descriptors of the Original Vegetation Datasets. Catchment = subject catchment, Match = Catchment the subject catchment is deemed a closest match with, Category = spatio-temporal category of the gauged catchment as established by previous research<sup>30</sup>. Gauged catchments are shown in the top plates, the ungauged catchments are shown in the bottom 2 plates and arranged chronologically from north to south. Visualisation of this data shows that some catchments have combinations of similar feature deviations to gauged catchments, and others are unique.

satisfactory to very good performance (Table 1). Meanwhile, training datasets that grouped all gauged catchments together only met satisfactory performance criteria for the scenario that included all spatial data variables (i.e., Ecounit, Land use and Original Vegetation (EULUOV)) (Table 1).

Performance criteria for the control (i.e., flow only) scenario also varied where the dataset was first discriminated to spatio-temporal categories (Table 1). Simulation results for the control scenario trained on non-discriminated datasets failed performance criteria (NSE = 0.130), while the opposite was the case for training datasets discriminated by spatio-temporal category (NSE = 0.846 and 0.947 for Category 2 and 3 respectively). The NSE for Category 1 catchment Control worsened after discrimination, however, the  $R^2$  value improved to 0.686 compared to 0.39 for the non-discriminated counterpart. Benefits of including spatial data in datasets were reduced after pre-discriminating to spatio-temporal regime (Table 1). Benefit losses include a lack of independence from the control scenario, as measured by Kruskal Wallis test for independence ( $p = 0.483\text{--}0.981$ ), where spatial data was omitted (Supplementary Material Table ST1).

Grouping datasets by respective catchment categories, identified in our previous spatio-temporal study<sup>30</sup>, prior to loading to the DIN simulator resulted in improved performance criteria ( $R^2 = 0.984$  for Category 3, RMSE = 0.02382 for Category 1). Interestingly, for Category 2 flow datasets the control scenarios, which did not contain spatial variables achieved superior performance for MSE,  $R^2$ , NSE and Wilmott's d compared to the other Category 2 scenarios that did include information on spatial variables. In contrast, the Original Vegetation scenario discriminated to Category 3 records had the smallest pde score meaning that the inclusion of Original Vegetation variables improved the ability of the DIN simulator under extremes in the data for Category 3 catchments (Table 1).

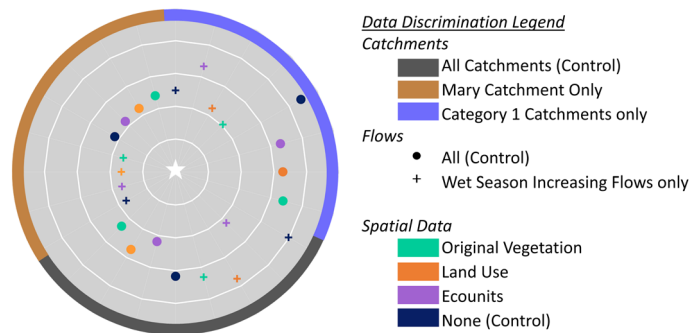
### Classifying ungauged to gauged catchments: variable independence vs ANN-PR

While the ANN-PR approach matched all ungauged catchments to a gauged counterpart, the XAI-SHAP variable independence approach using relative variable distributions was unable to match 17 catchments. Catchment matches using OV dataset for XAI-SHAP landform and floristic structure, most closely aligned to the ANN-PR catchment matches using the EU dataset. Matching using only the top 10% of deviated features using XAI-SHAP variable independence approach changed the catchment matches compared to the EU dataset using ANN-PR, where all variables are considered, but retained matches generally within the same category (Fig. 1).

### Verification of catchment classification for DIN similarities

Both XAI-SHAP Variable Independence and ANN-PR techniques for catchment classification matched pseudo-ungauged Herbert to the Gauged Mary (Figs. 1 and 2) and identified it as a Category 1 catchment. Only Mary catchment training data scenarios achieved a satisfactory performance metric i.e. NSE > 0.5 (Supplementary Material Table ST2). The greatest performance criteria overall collectively clustered towards datasets discriminated to Mary and Category 1 flows only (Fig. 3 and Supplementary Material Table ST2).

Datasets first discriminated by the classified catchment resulted in the best overall performance, with further discrimination to flow regime improving results. Training datasets discriminated by spatio-temporal flows, also performed better where they were also discriminated to the flow regime. This is consistent with findings during the development of the ANN-WQ simulator where a significant difference was noted for training datasets discriminated by the category flow regime. Where catchment specific or catchment category classification was not included, performance improved the most for training datasets that included Ecounit spatial data compared to the control which did not include spatial data. The worst performing scenario was the control grouped to Category 1 catchments only, followed by the control for all gauged catchments that included no spatial data, but was discriminated by flow regime. In comparison to control scenarios, differences in the performance criteria for scenarios that include spatial data diminished for catchments trained only to Mary gauged data. This



**Figure 3.** Additive Performance Criteria for pseudo-ungauged catchment DIN simulations. Dimensionless graph shows the additive scores for the best MSE,  $R^2$ , NSE, d, RMSE for each data discrimination scenario evaluated in the case study. Zero is the centrally located white star and represents the target/observed data. Discrimination of data to spatio-temporal regime (i.e., shown as crosses) and to the matched catchment improves results the most. Inclusion of Ecounit data improves simulation performance where training datasets are not discriminated to classified catchment or category.

suggests the benefit of adding spatial data reduced as the flow regime was refined to the catchment with the closest similarity to Herbert.

Training data discriminated to the individually matched catchment, Mary, and discriminated to wet season flows achieved the best performing DIN simulations ( $R^2 = 0.80$ ,  $NSE = 0.62$ ,  $d = 0.85$  respectively). Visualisation of simulated vs true data demonstrates that these scenarios' pre-discriminated spatio-temporal flows result in simulations that include all the peaks in the observed dataset. On the other hand, training data discriminated to include all catchments in the corresponding Category 1, but using all flow and season records, with no spatial data failed to simulate half the peaks (Supplementary Material Fig. SF2). While simulated peaks were under estimated in all cases, a review of the raw data identified that the maximum nitrogen concentration in the dataset for Herbert Catchment was 1.8105 mg/L, which is the highest historical record, plus two additional peaks ranging between 1.320 mg/L and 1.694 mg/L. Maximum concentration for Mary was a smaller with a once off observed peak of 1.243 mg/L during unusual weather conditions of end 2012 start 2013<sup>38,39</sup> with remaining peaks in the dataset not exceeding 0.605 mg/L.

## Discussion

### Overview

Our research uniquely evaluates the classification potential for all ungauged catchments flowing to the Great Barrier Reef, based on proxy data for spatio-temporal drivers of Dissolved Inorganic Nitrogen (DIN). We adopt an explainable AI approach referred to as XAI-SHAP to provide a deeper understanding of the modelled classification results. In accordance with earlier research works, our satisfactory performance metrics show classification of the pseudo-ungauged area to the most similar gauged ones is validated and works well where data for proxy drivers of DIN are included because they facilitate grouping of catchments by the DIN regime. Evaluation of DIN simulation performances using transfer learning in an Artificial Neural network environment allowed us to demonstrate the variability in DIN patterns depending on the spatio-temporal regime of the ungauged catchments, as exposed by original vegetation data. Additionally, the XAI-SHAP method allowed for ungauged catchments with insufficient similarity to the gauged ones to be identified, regardless of being classified by brute force using ANN-PR techniques.

### Dataset complexity and consistency

Development and verification of the ANN-WQ simulator to establish DIN response similarities in datasets between pseudo-ungauged catchments with the gauged ones found dataset complexity and representative flow patterns were influential. This highlighted caution in direct application without prior understanding of the DIN to flow dynamics of the catchment. Flatline simulations that resulted in the unsupervised scenario are a known symptom of inadequate complexity in the dataset<sup>40</sup>. Likely explanations include hidden neuron complexity was low in the development trials and relationships between flow, spatial data and DIN response was not adequately formed to facilitate simulations in the unsupervised scenario. The contrasting ability of Wet Tropics catchment datasets to overcome possible lack of dataset complexity in the training dataset is explained by the different DIN and flow dynamics in wet tropics catchments compared to the others<sup>27</sup>. We previously demonstrated that DIN remains elevated in retreating flows for Wet Tropics catchments only<sup>26</sup>. One explanation for the contrast with Wet Tropics catchments could be a more consistent relationship between flows and DIN releases throughout the hydrograph which the ANN-WQ simulator was trained to simulate for<sup>41,42</sup>. This phenomenon demonstrates that consistency of DIN to flow relationships influence the performance of defined algorithm based models developed to use transferred data.

### Training dataset influence

For catchments with inconsistent DIN to flow relationships, our results found training data arrangements that group catchments using prior knowledge of spatio-temporal similarities, i.e. either by prior discrimination (discriminated to Category 1, Category 2 and Category 3 as informed by Original Vegetation deviation using XAI-SHAP), or within the model training datasets (non-discriminated but including all EULOV spatial variables which are identified as proxy drivers for DIN and used to inform XAI-SHAP) improved the performance. This approach to remove heteroskedasticity where seasonal differences for Nitrate are considered has already been shown to benefit model development<sup>42,43</sup>. The significant differences in performance criteria of DIN simulations ( $p = 0.003-0.045$ ) depending on the data discrimination for catchment categories suggests that DIN dynamics differ between those categories. This finding of significant variation in nitrogen regimes through the Great Barrier Reef catchments, as demonstrated by the ANN-WQ simulator training dataset predictive performance, regardless of anthropogenic influence is consistent with proceeding research<sup>34,44,45</sup>. Our research shows variability in DIN regimes is an influential consideration for data transfer purposes in water quality models. The improved performance criteria where information on proxy drivers of DIN was considered supports our application of original vegetation spatial datasets used in this study to discriminate differences in DIN regimes for each catchment<sup>30</sup>.

Spatio-temporal category differences for DIN simulation performance may be explained via the wide body of literature that demonstrate nitrogen is either flow or production limited<sup>44,46,47</sup>, and also influenced by connectivity to stream network<sup>48</sup>. The superior performance of the control scenario for Category 2, compared to Category 2 scenarios that included spatial data, indicate this category is flow limited and abundant in DIN. It is demonstrated that soils higher in total organic carbon, consistent with rainforest soils, have higher supplies of nitrogen created by the residual soil biology<sup>30,49,50</sup>. The abundance of nitrogen generation in the soils, coupled with abundant flows, in the Wet Tropics catchments can result in consistent nitrogen to flow patterns and is a logical explanation for the ability of the ANN-WQ simulator to generate results in the development trials, where catchments from other categories flatlined. This supports our previous suggestions<sup>30</sup> that the timing of data

collection is important in Category 1 and 3 catchments, while Category 2 catchments could classify regardless of the season or flow phase. While our research is not designed to interrogate reasons for drivers of DIN in each category per-se, this is one of many possible explanations for how categorising datasets by vegetation removes noise associated with different combinations of biotic responses in each location<sup>34,41,51</sup>. The findings, therefore, support the second hypothesis that prior grouping of catchments by categories of Original Vegetation, as a proxy for the DIN to flow regime, is a necessary first step for identifying catchments that share similar DIN patterns.

Training dataset discrimination and variable combinations separately influenced the performance of the ANN-WQ simulations of DIN for the pseudo ungauged catchment. The ANN-WQ simulator achieved the best DIN simulation performance metrics for the pseudo-ungauged catchment when trained on data only from the classified catchment and therefore highlights that data transfer with the classified catchment achieved the best results. Concurrently, discriminating the dataset by the respective flow regime of wet season increasing flows had a greater influence on simulation performance than inclusion of spatial data variables. In contrast, training datasets using data from multiple catchments from the same flow regime, i.e. Category 1, achieved equivalent performance only where the training data was first discriminated to increasing above average flow regime, hence removing heteroskedasticity of DIN in the retreating and below average flows. Both these findings are consistent with the ANN-WQ development phase and demonstrate that prior discrimination of the training dataset to flow regime reduces heteroskedasticity in DIN patterns to flow<sup>41,52</sup>. Once heteroskedasticity in the training dataset was removed, the influence of spatial variables as drivers to the DIN patterns became less relevant. Separately, the research also found that training the ANN-WQ simulator using data from all catchments improved where all EULUOV variables were included. This could be attributed to the ANN-WQ simulator discriminating datasets within the algorithms, as opposed to prior discrimination provided by classification, and further demonstrates the benefit of the spatial datasets to expose the drivers of the DIN patterns.

### ANN-PR vs XAI-SHAP classification

Catchments matched using ANN-PR were not always the same as the catchments recommended to be matched by the XAI-SHAP deviation approach for variable independence. One reason could be that only 10% of the most influential variables were considered in the XAI-SHAP approach, in contrast the less deviated variables contributed to the ANN-PR matches. Catchment classification informed by the match options in both ANN-PR and XAI-SHAP approaches provide foundational guidance to rationalise catchments to evaluate in future data transfer investigations or models for DIN simulations to the Great Barrier Reef<sup>53</sup>. Varied performance for each scenario trialled in the ANN-WQ simulator development phase demonstrated that training data from catchments with the most similar proxy drivers of DIN dynamics, was more suitable for data transfer compared to training data from all catchments lumped together<sup>54</sup>. This demonstrates that rationalising training data to the most similar responding catchments reduces heteroskedasticity in the training dataset and benefits DIN simulation accuracy for the classified catchment. XAI-SHAP provided insight to identify catchments grouped by known DIN to flow proxy drivers. While classification using nearest neighbour catchments has historically been supported for their influence towards flow similarities<sup>15,55</sup>, our finding demonstrates that catchments with the most similar drivers of DIN, in addition to flow, are not necessarily located as the neighbouring catchment, and are influential towards DIN simulation performance.

### Practical application

This study established Original Vegetation as a suitable proxy for DIN dynamics for the benefit of water quality modelling. Therefore, the 20 ungauged catchments that matched to gauged ones, based off Original Vegetation similarity have justification to receive data from the corresponding gauged catchment. The remaining 21 ungauged catchments had combinations of original vegetation unique from the gauged catchments and therefore did not support the hypothesis that they share similar DIN drivers with gauged catchments. Consequently, this study found that only 20 of the 41 ungauged catchments were suitable to consider for data transfer with existing gauged catchments for satisfactory water quality modelling purposes.

Of the ungauged catchments that failed to match to gauged ones, 4 groups shared unique combinations of deviated spatial variables. Deviated original vegetation floristic structure and landform descriptors shown by the XAI-SHAP deviations, were grassland, heathland and mangrove. These are all coastal ecosystems and differ from vineforest, open woodlands and forest shown to be proxy indicators of DIN dynamics for gauged catchments<sup>30</sup>. While data transfer from existing gauged catchments to the four coastal catchments is not supported by our study, our method can instead be used to inform where new water quality monitoring and gauging sites could have the greatest value to represent all DIN regimes<sup>12,47,56–58</sup>. New monitoring and gauging sites are recommended in each of the four coastal catchment groups to collect data representative of all DIN regimes, which could facilitate data transfer for modelled DIN predictions across all ungauged Great Barrier Reef catchments.

It is well known that performance of neural networks deteriorates when the unsupervised scenario includes extremes outside the range of the training dataset<sup>59</sup> and in our evaluation, models trained using Mary data were never exposed to high concentration peaks observed in the Herbert catchment. Limitation for simulating extremes not included in the training data could be addressed with differing model techniques, the ANN-WQ simulator was intended only as a coarse method to verify whether similarity in DIN drivers exists between catchments matched using the ANN-PR method, and this was demonstrated.

For the case study, the matched catchment was a Category 1 catchment. Collectively, Category 1 catchments showed the poorest performance in the ANN-WQ development phase. The fact the case study trial achieved satisfactory performance criteria for the poorest performing category in the development phase, it is expected that better results can be achieved for Category 2 and Category 3 catchments which have less heteroskedastic DIN to flow relationships. Further studies to refine the ANN-WQ simulator performance, along with a full



comparison of all training dataset options, i.e. discrimination of data to Category 2 and 3 flows to evaluate difference in performance and facilitate year round classification is recommended. Regardless of these limitations, we encourage results from this study to be applied in established models that will benefit from data transfer from the most similar catchments for purpose of DIN modelling, and intentionally developed for superior performance<sup>43</sup>.

## Conclusions

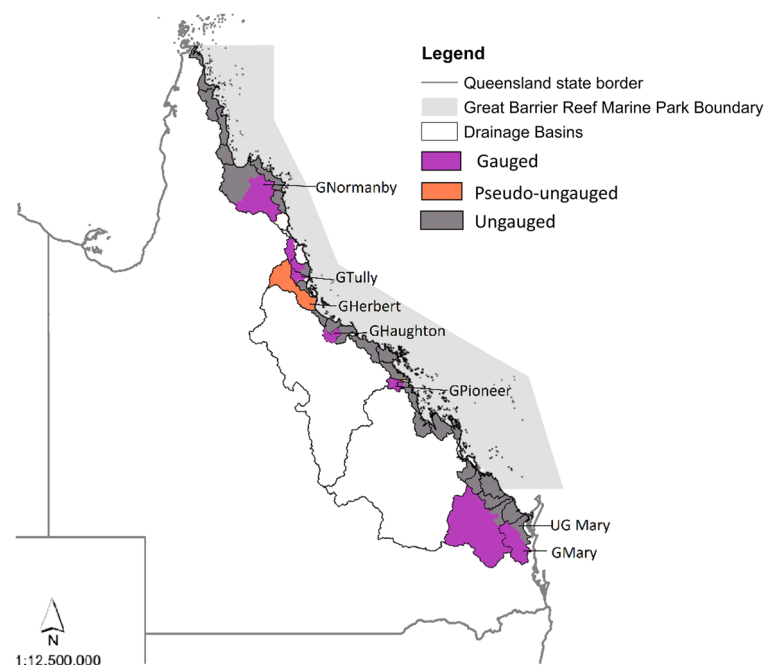
This study matched all ungauged catchments that drain to the Great Barrier Reef to the gauged ones using ANN-PR coupled with Land use and Original Vegetation datasets. While ANN-PR enabled matching using proxy datasets for drivers of DIN, XAI-SHAP method explained similarities between catchments based on feature deviations as well as concurrently allowing grouping of catchments to known spatio-temporal categories. Prior knowledge of spatio-temporal DIN response categories within training datasets improved performance of the ANN-WQ simulator developed to verify catchment matches.

While all catchments matched to a gauged one using ANN-PR, consistent with hypothesis 1, the additional interrogation by XAI-SHAP deviations found 17 catchments did not share deviated feature similarity with a spatio-temporal category. The XAI-SHAP method instead provides justification to prioritise gauging and monitoring efforts in those unmatched catchments to better understand the spatial temporal dynamics of DIN in coastal areas that those unmatched catchments were located in. For the ungauged catchments that did match to gauged ones using the XAI-SHAP method, the subsequent ANN-WQ simulator development and case study to test the second hypothesis, found prior discrimination of data included in the training dataset, based on the spatio-temporal category of the ungauged catchment, improved performance of the ANN-WQ simulator in all scenarios tested. It was, however, an unexpected finding that, after the spatio-temporal discrimination by category was first applied, inclusion of Original Vegetation, Ecounit or Land Use variables had insignificant influence on results. Findings that emerged throughout this study therefore built nuance to our expected hypothesis 3 whereby although a trained ANN-WQ simulator successfully simulated DIN in the unsupervised scenario, it was the knowledge provided by original vegetation data to pre-process the training datasets into categories that mattered. Implications of these findings are that XAI-SHAP coupled with Original Vegetation data has demonstrated merit for customising catchment matching to the portion of water quality datasets most likely to share similar DIN to flow regimes between gauged and ungauged catchments.

## Methods

### Study area

This study includes all catchments that flow to the Great Barrier Reef, in north-eastern Australia. Each of those catchments is referred to herein as gauged, ungauged and pseudo-ungauged as shown in Fig. 4. Respective gauging allocation, sampling frequency for DIN, and flow data availability for each of the catchments are provided in Supplementary Material Table ST3.



**Figure 4.** Study Area. G preceding catchment name infers a true gauged catchment. UG preceding catchment name infers ungauged catchment. Maps created by author using ArcMap 10.8.1, ungauged catchments<sup>10</sup> supplied, Drainage Basins<sup>37</sup> licenced under a Creative Commons—Attribution 3.0 Australia licence (CC BY 3.0 AU). © State of Queensland (Department of Environment and Science) 2023.

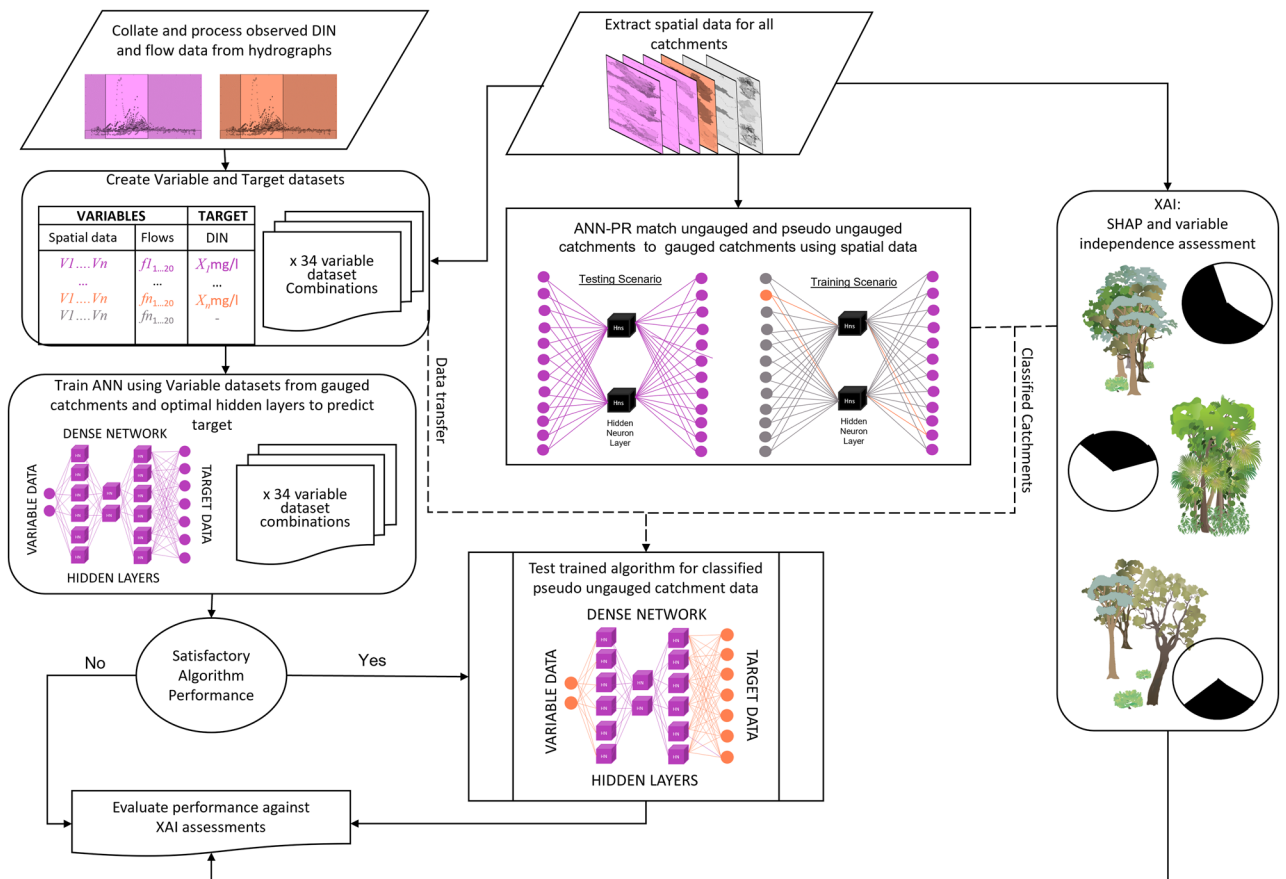
### Study concept

The objective of this study is to establish whether patterns in the flow and spatial variable datasets contain sufficient information to simulate Dissolved Inorganic Nitrogen (DIN), and whether forecasting capabilities can extend to new catchments, referred to in this study as pseudo-ungauged. Because the influence of every variable input and their interrelationships to overall DIN response are unknown a priori, a dense fully connected Artificial Neural Network (ANN) algorithm was developed to trial the proof of concept approach. Algorithms were trained for a number of dataset arrangements and their performance metrics were compared to quantify the viability of the novel forecasting/data transfer concept within the Artificial Intelligence modelling environment. A workflow conceptualising the research approach is shown in Fig. 5 below.

### Study dataset

Observed water quality data and flow records for those gauged and pseudo-ungauged catchments are from locations listed in Supplementary Material Table ST3. This data was sourced from Queensland State Government and was cleaned, transformed and flows arranged as detailed in our foundational research<sup>26</sup>. The spatial extent of gauged areas for catchments evaluated in this study are consistent with Khan et al.<sup>13</sup>. DIN records were collected at irregular frequencies depending on flows for each gauged catchment as detailed in Supplementary Material Table ST3. To overcome a large number of NaN values within a time series arrangement for the dataset, daily average stream and baseflows for 90 days preceding each DIN record were allocated as 90 separate column variables each on the same dataset row position as the corresponding DIN record as 1 day prior, 2 day prior...90 day prior. 90 days prior flows capture a full temperate climate season preceding each DIN record and were shown by cross correlations to be sufficient to capture residual information<sup>60</sup>. The water quality and flow datasets were duplicated then partitioned as outlined in O’Sullivan et al.<sup>30</sup> for wet season/increasing flows, dry season/retreating flows, and all flows/seasons to capture spatio-temporal influences.

Spatial data for all gauged and ungauged portions of catchments in the study area, were extracted from Queensland Government Q-Spatial mapping platform, as per the methods described in O’Sullivan et al.<sup>26</sup>. The three separate spatial datasets were created as the proxy drivers of DIN. These included: Land use to represent



**Figure 5.** Conceptual framework of research. This framework shows data preparation for classification, process for simulating DIN for pseudo ungauged catchments, and relationship to XAI evaluation of results. Data relating to gauged catchments is represented in purple, orange represents pseudo-ungauged catchments, and ungauged catchments are represented by grey. Dashed lines show source and destination of data transfer for pseudo-ungauged catchment in the research. Vegetation images adapted from the Integration and Application Network licenced under Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

human biotic influence which included 6 variables<sup>61</sup>, Original Vegetation consisting of 38 variables<sup>31</sup> intended as a parsimonious biotic response proxy for natural DIN responses across the catchment<sup>26,62</sup>, and Ecounit which was created via a combination of Land use and Original Vegetation and resulted in 179 variables. The area of variables for each catchment was established via clipping the spatial datasets to the catchment boundaries and extracting corresponding data tables from ArcGIS. The area of the gauged catchments extended only to the gauged monitoring point, the ungauged portion was created as a sub-catchment polygon for all areas that drain to the catchment's waterway downstream of the gauged monitoring point, or for fully ungauged catchments. For each catchment, the spatial dataset was duplicated to match the number of data rows to the same number of DIN records in each catchment dataset. For the pseudo-ungauged datasets, the number of spatial dataset rows were duplicated to match to the number of daily average flow records available.

A master dataset was created by joining the preceding flow and spatial dataset to create the training variables dataset, and the corresponding DIN data allocated as the target dataset. All data in each dataset was then normalised. Scenario datasets were then created by extracting subsets of data from the master dataset as detailed in Table 2.

### Classifying gauged catchments to ungauged and pseudo-ungauged catchments

The novel aspect of this research is establishing whether pseudo-ungauged, and ungauged catchments share spatial data similarities suitable for classifying to gauged catchment classifiers, and for water quality classification data transfer purposes. Our previous studies used ANN-PR to classify only the gauged catchments together using the same spatial variables used in this study<sup>26,30</sup>. XAI evaluations of those datasets provide explainability to the

Dataset type	Purpose	Catchments included	Dataset reference	Variables (discriminated by):				Target data	
				LU spatial variables	OV spatial variables	EU spatial variables	90 day prior flows	Corresponding DIN	Corresponding catchment
1. Spatial Classification Training Datasets	Training algorithms to match spatial data to corresponding gauged catchment	All gauged	1.GCAFULU	✓	–	–	–	–	✓
			1.GCAFAOV	–	✓	–	–	–	✓
			1.GCAFAEU	–	–	✓	–	–	✓
			1.GCAFAEULUOV	✓	✓	✓	–	–	✓
2. Spatial Classification Testing Datasets	Classify pseudo-ungauged and ungauged Catchments to gauged catchment	Individual ungauged/pseudo-ungauged	2.CAFALU	✓	–	–	–	–	Hidden
			2.CAFAOV	–	✓	–	–	–	Hidden
			2.CAFAEU	–	–	✓	–	–	Hidden
			2.CAFAEULUOV	✓	✓	✓	–	–	Hidden
3. ANN-WQ simulator development- training datasets	Establish whether recognisable patterns exist in the datasets to forecast DIN	All gauged together (non-discriminated)	3.GCAFULU	✓	–	–	✓	✓	–
			3.GCAFAOV	–	✓	–	✓	✓	–
			3.GCAFAEU	–	–	✓	✓	✓	–
			3.GCAFAEULUOV	✓	✓	✓	✓	✓	–
		All gauged—individual (Discriminated by catchment)	3.GC <sub>i1...in</sub> FALU	✓	–	–	✓	✓	–
			3.GC <sub>i1...in</sub> FAOV	–	✓	–	✓	✓	–
			3.GC <sub>i1...in</sub> FAEU	–	–	✓	✓	✓	–
			3.GC <sub>i1...in</sub> FAEULUOV	✓	✓	✓	✓	✓	–
		Gauged Catchments grouped by Spatio-Temporal Category (Discriminated by Category 1, Category 2 or Category 3)	3.GC <sub>1...3</sub> F <sub>1...3</sub> LU	✓	–	–	✓	✓	–
			3.GC <sub>1...3</sub> F <sub>1...3</sub> OV	–	✓	–	✓	✓	–
			3.GC <sub>1...3</sub> F <sub>1...3</sub> EU	–	–	✓	✓	✓	–
			3.GC <sub>1...3</sub> F <sub>1...3</sub> EU-LUOV	✓	✓	✓	✓	✓	–
4. ANN-WQ Simulator-Trial Datasets	Evaluate suitability of matched dataset for data transfer to pseudo-ungauged catchment for DIN simulation purposes	Individual gauged and pseudo-ungauged catchment	4.C <sub>i1...in</sub> FALU	✓	–	–	✓	Hidden	–
			4.C <sub>i1...in</sub> FAOV	–	✓	–	✓	Hidden	–
			4.C <sub>i1...in</sub> FAEU	–	–	✓	✓	Hidden	–
			4.C <sub>i1...in</sub> FAEULUOV	✓	✓	✓	✓	Hidden	–
		Gauged and pseudo-ungauged catchments grouped by categories	4.C <sub>1...3</sub> F <sub>1...3</sub> LU	–	✓	–	✓	Hidden	–
			4.C <sub>1...3</sub> F <sub>1...3</sub> OV	–	✓	–	✓	Hidden	–
			4.C <sub>1...3</sub> F <sub>1...3</sub> EU	–	✓	–	✓	Hidden	–
			4.GC <sub>1...3</sub> F <sub>1...3</sub> EU-LUOV	✓	✓	✓	✓	Hidden	–

**Table 2.** Summary of data included in scenario datasets. Variations in each dataset intended to evaluate the influence of spatial data or flow variable toward DIN response. G= Gauged, C Catchment, F Flows, A All, LU LandUse, OV Original Vegetation, EU Ecounits C<sub>i1...in</sub> reference to individual catchments.

corroborating ANN-PR results for both spatial data and water quality classification<sup>30</sup>. Here, we explore, for the first time, extending that classification approach beyond the gauged portion of the study area to classify catchments of the ungauged and pseudo-ungauged areas to gauged catchments. The method is therefore extending the ANN-PR approaches of our previous studies to now evaluate which gauged catchments the pseudo-ungauged and ungauged catchments classify to, and evaluate whether XAI explainability applies to the ANN-PR matches. To accomplish this, we apply a combination of the ANN-PR approach used in our previous studies and XAI explainability coupled with SHAP<sup>36</sup> (XAI-SHAP) to evaluate the similarities between the catchments. This allows classification of catchments that have not been gauged, based on the similarities between the gauged catchments, and provide a better understanding of the underlying similarities between the catchments. Importantly, inclusion of the XAI-SHAP method demonstrates whether the sufficient underlying similarity is likely to exist between the proxy drivers of DIN in the gauged and ungauged catchments for the purpose of data transfer.

### Spatial classification using ANN-PR

This step used a similar approach explained in detail in our previous studies, however this time we trained the ANN-PR tool on all 11 gauged catchments, and introduced the spatial variable data for the ungauged catchments in the unsupervised environment, to force a match to one of the 11 gauged catchments. A 100-fold duplicate of each spatial variable in each gauged catchment was used to estimate the percentage match between the ungauged catchments and the gauged catchments. We then trained the ANN-PR classification tool in a supervised environment by applying the gauged catchment classification training datasets to standard codes extracted from “MATLAB 2020a (The MathWorks Inc., 2020) Deep Learning toolbox (Fig. 4). The code used is a two-layer feed-forward network, with sigmoid transfer function in the hidden layer, and softmax transfer function in the output layer (The MathWorks Inc. 2020)”<sup>63</sup>. For the spatial datasets, heuristics and previous knowledge for the gauged data spatial dataset meant that an architecture of 3 hidden neurons were used to set the classification training architecture for this model. Data were split within the coding architecture to 70% for network training, 15% network validation and 15% network testing. In the training phase, the network is designed to match spatial data variables for each row in the dataset to one of the 11 gauged catchment categories the spatial data is sourced from. The network architecture is set such that training continues towards minimisation of cross entropy and stops once mean square error elevates above its minimum pivot point at which point the ANN-PR algorithm achieves optimal performance<sup>64</sup>. Optimal performance is for each of the 100 replicates of spatial data to allocate to the catchment category the data belonged to in the validation and testing phase.

Testing datasets were separately introduced in an unsupervised environment to the optimised classification algorithm trained to match spatial data to only one of the 11 gauged catchments. Spatial variables for each ungauged or pseudo-ungauged catchment were duplicated 100 times so that the catchment the ungauged or gauged spatial dataset was classified to was based on 100 replicates. The algorithm forces each of the 100 rows of spatial data variables the ungauged or pseudo-ungauged catchment to match to one of the 11 classifiers in the trained environment. This approach was repeated for the Land use, Original Vegetation and Ecounit spatial datasets for all 41 ungauged and pseudo-ungauged catchments. The gauged catchment with allocations of more than half the records for each gauged or ungauged catchment was deemed classified for the respective dataset.

### Identifying variable feature independence in both gauged and ungauged catchments

The purpose of XAI, is to deduce the combination of variables most likely to have resulted in the classification between two catchments. To verify that the forced matches between gauged and ungauged catchments using ANN-PR were explainable, we therefore extend the additive deviation approach from previous work<sup>30</sup> to spatial variables for all catchments, shown in Eq. (1), and graphed the top 10%<sup>30</sup>.

$$D_s = A_s - A_v \quad (1)$$

where: D: deviation of spatial dataset variable. A: proportional area of variable ( $A = \text{area of variable} / \text{total catchment area}$ ), S: subject variable,  $\forall$ : all dataset variables excluding S.

Variables in the top 10% deviated from the mean were then graphed and visually compared for similarities between the deviated variables for gauged and ungauged catchments sharing similar combinations of deviations were categorised together. Because the Original Vegetation dataset had previously been shown to explain the ANN-PR matches between the gauged catchments<sup>30</sup> it was used directly in this study. Geology and landform has also been demonstrated as a fundamental driver of nitrate in hydrological processes<sup>48</sup>, therefore in this study we also further scrutinised influences of the original vegetation dataset by breaking each variable down into its separate landform type and floristic structure descriptor as described by the data authority<sup>31</sup> to better visualise hydrological drivers of results.

### Training ANN to forecast DIN in a supervised environment

An Artificial Neural Network water quality (ANN-WQ) simulator was developed to facilitate a rapid assessment of the similarity of matched catchments for DIN. The ANN-WQ simulator was intended for rapid comparison purposes only, and therefore method optimisation was outside the catchment classification scope intended for this research. Similarity between catchments was evaluated by the comparative accuracy of DIN simulations generated for catchments depending on the dataset scenario included in the ANN-WQ simulator training phase.

For each gauged catchment dataset, a Dense Deep Learning feed forward network was created in Matlab. The dense fully connected learning approach was selected to facilitate for all data relationships to be considered, to maximise the pattern recognition ability within the dataset, timesteps of data are still captured in variables as the corresponding time-date number. This architecture was resource intensive and therefore a ReLU hidden layer

Performance metric	Equation	Satisfactory criteria
Correlation Coefficient (R <sup>2</sup> )	$R^2 = \left( \frac{\sum_{i=1}^N (Y_i^{obs} - Y_{mean}^{obs})(Y_i^{sim} - Y_{mean}^{sim})}{\sqrt{\sum_{i=1}^N (Y_i^{obs} - Y_{mean}^{obs})^2} \sqrt{\sum_{i=1}^N (Y_i^{sim} - Y_{mean}^{sim})^2}} \right)^2$ Equation 2 <sup>69</sup>	> 0.5
Nash–Sutcliffe coefficient (NSE)	$NSE = 1 - \left[ \frac{\sum_{i=1}^N (Y_i^{obs} - Y_i^{sim})^2}{\sum_{i=1}^N (Y_i^{obs} - Y_{mean}^{obs})^2} \right]$ Equation 3 <sup>54,70</sup>	> 0.5
Willmotts index (d)	$d = 1 - \left[ \frac{\sum_{i=1}^N (Y_i^{obs} - Y_i^{sim})^2}{\sum_{i=1}^N ( Y_i^{sim} - Y_{mean}^{obs}  +  Y_i^{obs} - Y_{mean}^{sim} )^2} \right]$ Equation 4 <sup>71</sup>	> 0.5
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i^{sim} - Y_i^{obs})^2}$ Equation 5 <sup>69,72</sup>	Lowest
Peak percentage deviation (pde)	$100 \sum_{i=1}^N \frac{1 - Y_{max}^{sim}}{Y_{max}^{obs}}$ Equation 6 <sup>73</sup>	< ± 25
Mean absolute error	$MAE = \frac{1}{N} \sum_{i=1}^N  Y_i^{sim} - Y_i^{obs} $ Equation 7 <sup>72</sup>	Lowest

**Table 3.** Performance Metrics and nominated criteria for ANN\_WQ simulation scenarios. Where: N = number, i = iteration,  $Y_i^{obs}$  = Observed data,  $Y_i^{sim}$  = target data from model simulation,  $Y_{mean}^{sim}$  = mean of the simulation,  $Y_{mean}^{obs}$  = mean of observed,  $Y_{max}^{obs}$  = maximum of observed,  $Y_{max}^{sim}$  = maximum of simulation.

activation was included due to its superior ability to deal with weights and bias over large intensity variations, as could be expected in the dataset<sup>65,66</sup>.

Training datasets involved a data set split of 80% Training, 10% Verification and 10% Testing. Development of the Dense Deep Learning feed forward network began with a trial and error phase to scope for functionality at the default hidden neurons (< 10). To overcome inadequate complexity and dimensionality within datasets, trials of 1 to 1000 hidden layers were then undertaken for each dataset to identify the best performing hidden layer network suited to the training dataset<sup>67</sup>. Trialling up to 1000 hidden layers on big data creates heavy computing demands, therefore, Adam optimiser was selected for its minimal memory usage benefits whilst also addressing sparse gradients and non-stationary objectives<sup>68</sup>. The model performance metrics comprised of RMSE, MSE, Nash Sutcliffe Efficiency, Peak Deviation and Correlation as R<sup>2</sup> were recorded for each of the hidden layer trials, and the algorithm with the best performance metrics evaluated for the optimal hidden neuron and for pass or fail of satisfactory performance criteria. The performance metrics equations in Table 3 identify the corresponding satisfactory performance criteria for each. For this research, the ANN-WQ simulator was used to validate whether DIN patterns were detectable. Therefore, performance criteria that identified whether the results were satisfactory or not as nominated in Table 3 were selected to remain consistent with satisfactory performance criteria for water models published elsewhere<sup>69–73</sup>.

To compare model DIN forecasts against observed DIN forecasts, the algorithm was rerun with the optimal number of hidden layers for every dataset. We normalized all data for graphing so we could compare forecasting potential across the different datasets.

### DIN forecasting potential for classified pseudo-ungauged catchments

The trained algorithms that met minimum satisfactory performance criteria as well as demonstrating a simulation ability in the supervised environment were then used in an unsupervised environment to simulate DIN for their respective classified pseudo-ungauged catchment datasets based on flow inputs. For the data available, only the pseudo-ungauged Herbert was suitable for evaluation for the study and is evaluated as a case study within this article as proof of concept. For this study, scenario datasets evaluated included the ANN-WQ simulation results for the pseudo-ungauged catchment trained on the matched gauged catchment, ANN-WQ simulator trained using all gauged catchment data, and ANN-WQ simulator trained using data from the matching spatio-temporal category. Performance metrics for each scenario were then collated and visualised in a dart plot. To create the dart plot, performance metrics were adjusted using Eq. (8) to make zero the target score. This equation has not been scaled for the impact each performance metric has towards the accuracy of the model, but is developed here for rapid comparison of overall scenario performance.

$$\cup_{PC_k} = \sum_{k \in \{PC_1, \dots, PC_n\}} k, \tag{8}$$

where: PC = Performance Criteria, K = unsupervised portion of ANN-WQ simulation scenario, 1...n = a performance criteria adjusted to make zero target i.e. {R<sup>2</sup>, NSE, d, RMSE, MAE, Pdv}.

Where:

$$R^2_c = 1 - R^2.$$

$$NSE_c = 1 - NSE.$$

$$pde_c = (pde \text{ if } pde > 0 \mid - pde \text{ if } Pdv < 0) \times 0.01$$

## Data availability

The datasets analysed during the current study are available from the corresponding author on reasonable request, as well as in raw form from the following public sources: Observed water quality and flow records<sup>37,74</sup>—Queensland Government Water Monitoring Information Portal: <https://water-monitoring.information.qld.gov.au/>. Original Vegetation<sup>31</sup>—Pre-clearing broad vegetation groups—Queensland (v4): [http://qldspatial.information.qld.gov.au/catalogue/custom/search.page?q=%22Pre-clearing broad vegetation groups - Queensland%22](http://qldspatial.information.qld.gov.au/catalogue/custom/search.page?q=%22Pre-clearing%20broad%20vegetation%20groups%20-%20Queensland%22). Land Use<sup>61</sup>—Land use mapping—1999 to 2017—Queensland <http://www.qld.gov.au/environment/land/vegetation/mapping/qlump/>.

Received: 2 June 2023; Accepted: 17 October 2023

Published online: 24 October 2023

## References

- Creighton, C., Waterhouse, J., Day, J. C. & Brodie, J. Criteria for effective regional scale catchment to reef management: A case study of Australia's Great Barrier Reef. *Mar. Pollut. Bull.* **173**, 112882. <https://doi.org/10.1016/j.marpolbul.2021.112882> (2021).
- Gupta, A. & Govindaraju, R. S. Propagation of structural uncertainty in watershed hydrologic models. *J. Hydrol.* **575**, 66–81. <https://doi.org/10.1016/j.jhydrol.2019.05.026> (2019).
- Vu, D. T., Dang, T. D., Pianosi, F. & Galelli, S. Calibrating macro-scale hydrological models in poorly gauged and heavily regulated basins. *Hydrol. Earth Syst. Sci. Discuss.* <https://doi.org/10.5194/hess-2023-35> (2023).
- Garna, R. K., Fuka, D. R., Faulkner, J. W., Collick, A. S. & Easton, Z. M. Watershed model parameter estimation in low data environments. *J. Hydrol. Reg. Stud.* **45**, 101306. <https://doi.org/10.1016/j.ejrh.2022.101306> (2023).
- Tegegne, G. & Kim, Y. O. Modelling ungauged catchments using the catchment runoff response similarity. *J. Hydrol.* **564**, 452–466. <https://doi.org/10.1016/j.jhydrol.2018.07.042> (2018).
- Hrachowitz, M. *et al.* A decade of predictions in ungauged basins (PUB)—A review. *Hydrol. Sci. J.* **58**(6), 1198–1255. <https://doi.org/10.1080/02626667.2013.803183> (2013).
- Jahanshahi, A., Patil, S. D. & Goharian, E. Identifying most relevant controls on catchment hydrological similarity using model transferability—A comprehensive study in Iran. *J. Hydrol.* **612**, 128193. <https://doi.org/10.1016/j.jhydrol.2022.128193> (2022).
- UN General Assembly (2015), Transforming our world: the 2030 Agenda for Sustainable Development, 21 October 2015, A/RES/70/1, United Nations, New York. Available at: <https://sdgs.un.org/publications/transforming-our-world-2030-agenda-sustainable-development-17981> [Accessed 16 March 2021].
- Ani, C. J., Smithers, S. G., Lewis, S., Baird, M. & Robson, B. eReefs modelling suggests Trichodesmium may be a major nitrogen source in the Great Barrier Reef. *Estuarine Coast. Shelf Sci.* <https://doi.org/10.1016/j.ecss.2023.108306> (2023).
- Wells, S. C., Cole, S. J., Moore, R. J., Khan, U., Hapuarachchi, P., Hasan, M. M., Gamage, N., Bari, M.A., & Tuteja, N. K. (2019). Distributed hydrological modelling for forecasting water discharges from the land area draining to the Great Barrier Reef coastline. In *Geophysical Research Abstracts*. Vol. 21.
- Herr, A. & Kuhnert, P. M. Assessment of uncertainty in Great Barrier Reef catchment models. *Water Sci. Technol.* **56**(1), 181–188. <https://doi.org/10.2166/wst.2007.450> (2007).
- Kuhnert, P. M. *et al.* Making management decisions in the face of uncertainty: A case study using the Burdekin catchment in the Great Barrier Reef. *Mar. Freshw. Res.* **69**(8), 1187–1200. <https://doi.org/10.1071/MF17237> (2018).
- Montanari, A. *et al.* “Panta Rhei—Everything flows”: Change in hydrology and society—the IAHS scientific decade 2013–2022. *Hydrol. Sci. J.* **58**(6), 1256–1275. <https://doi.org/10.1080/02626667.2013.809088> (2013).
- Khan, U. *et al.* Development of catchment water quality models within a realtime status and forecast system for the Great Barrier Reef. *Environ. Modell. Softw.* **132**, 104790. <https://doi.org/10.1016/j.envsoft.2020.104790> (2020).
- Liu, S. *et al.* Characterisation of spatial variability in water quality in the Great Barrier Reef catchments using multivariate statistical analysis. *Mar. Pollut. Bull.* **137**, 137–151. <https://doi.org/10.1016/j.marpolbul.2018.10.019> (2018).
- Oudin, L., Andréassian, V., Perrin, C., Michel, C. & Le Moine, N. Spatial proximity, physical similarity, regression and ungauged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resour. Res.* <https://doi.org/10.1029/2007WR006240> (2008).
- Guo, Y., Zhang, Y., Zhang, L. & Wang, Z. Regionalization of hydrological modelling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdiscipl. Rev. Water* **8**(1), e1487. <https://doi.org/10.1002/wat2.1487> (2021).
- Yaseen, Z. M., Sulaiman, S. O., Deo, R. C. & Chau, K.-W. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *J. Hydrol.* **569**, 387–408. <https://doi.org/10.1016/j.jhydrol.2018.11.069> (2019).
- Merz, R., Tarasova, L. & Basso, S. Parameter's controls of distributed catchment models—How much information is in conventional catchment descriptors?. *Water Resour. Res.* <https://doi.org/10.1029/2019WR026008> (2020).
- Maestre, A., El-Sheikh, E., Williamson, D. & Ward, A. A machine learning tool for weighted regressions in time, discharge, and season. *Int. J. Adv. Comput. Sci. Appl.* <https://doi.org/10.14569/IJACSA.2014.050314> (2014).
- Noori, N., Kalin, L. & Isik, S. Water quality prediction using SWAT-ANN coupled approach. *J. Hydrol.* **590**, 125220 (2020).
- Chen, S. *et al.* Physics-guided machine learning from simulated data with different physical parameters. *Knowl. Inform. Syst.* <https://doi.org/10.1007/s10115-023-01864-z> (2023).
- Appling, A. P., Oliver, S. K., Read, J. S., Sadler, J. M. & Zwart, J. *Machine Learning for Understanding Inland Water Quantity, Quality, and Ecology* (Elsevier, 2022). <https://doi.org/10.1016/B978-0-12-819166-8.00121-3>.
- Nash, J. E. & Sutcliffe, J. V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **10**(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6) (1970).
- Liu, S. *et al.* A multi-model approach to assessing the impacts of catchment characteristics on spatial water quality in the Great Barrier Reef catchments. *Environ. Pollut.* **288**, 117337. <https://doi.org/10.1016/j.envpol.2021.117337> (2021).
- O'Sullivan, C. M. *et al.* Classification of catchments for nitrogen using Artificial Neural Network Pattern Recognition and spatial data. *Sci. Total Environ.* **809**, 151139. <https://doi.org/10.1016/j.scitotenv.2021.151139> (2022).
- Calvert, J., McTaggart, A., Carvalhais, L. C., Drenth, A. & Shivas, R. Communities of tropical soil fungi differ between burned and unburned forest, with corresponding changes in plant community composition, litter and soil chemistry. *bioRxiv* <https://doi.org/10.1101/2021.08.22.457293> (2021).
- Sepp, S. K. *et al.* Global diversity and distribution of nitrogen-fixing bacteria in the soil. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2023.1100235> (2023).
- Lintern, A. *et al.* Key factors influencing differences in stream water quality across space. *Wiley Interdiscipl. Rev. Water* **5**(1), e1260. <https://doi.org/10.1002/wat2.1260> (2018).
- O'Sullivan, C. M., Ghahramani, A., Deo, R. C. & Pembleton, K. G. Pattern recognition describing spatio-temporal drivers of catchment classification for water quality. *Sci. Total Environ.* **861**, 160240. <https://doi.org/10.1016/j.scitotenv.2022.160240> (2023).

31. Neldner, V. J., Niehus, R. E., Wilson, B. A., McDonald, W. J. F., Ford, A. J., & Accad, A. (2017). The vegetation of Queensland. Descriptions of broad vegetation groups. Version 3.0. Queensland herbarium, Department of Science, Information Technology and Innovation.
32. Wachter, S., Mittelstadt, B. & Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. J.L. Tech.* **31**, 841 (2017).
33. Wu, J. & Lu, J. Spatial scale effects of landscape metrics on stream water quality and their seasonal changes. *Water Res.* **191**, 116811. <https://doi.org/10.1016/j.watres.2021.116811> (2021).
34. Liu, S. *et al.* A Bayesian approach to understanding the key factors influencing temporal variability in stream water quality—a case study in the Great Barrier Reef catchments. *Hydrol. Earth Syst. Sci.* **25**(5), 2663–2683. <https://doi.org/10.5194/hess-25-2663-2021> (2021).
35. Wang, Y. N., Li, B. & Yang, G. Stream water quality optimized prediction based on human activity intensity and landscape metrics with regional heterogeneity in Taihu Basin, China. *Environ. Sci. Pollut. Res.* **30**(2), 4986–5004. <https://doi.org/10.1007/s11356-022-22536-5> (2023).
36. Shapley, L. S. (1953). A value for n-person games. Contributions to the Theory of Games II, Annals of Mathematical Studies, 28 (1953)
37. State of Queensland (Department of Environment and Science) (2018) Great Barrier Reef catchment and river basins <http://qldspatial.information.qld.gov.au/catalogue/custom/search.page?q=%22Great%20Barrier%20Reef%20catchment%20and%20river%20basins%22>.
38. Australian Bureau of Meteorology. (2013). Special climate statement 43—extreme heat in January 2013.
39. Australian Bureau of Meteorology. (2013). Special climate statement 44—extreme rainfall and flooding in coastal Queensland and New South Wales.
40. Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249–256). JMLR Workshop and Conference Proceedings.
41. Hirsch, R. M. Large biases in regression-based constituent flux estimates: Causes and diagnostic tools. *JAWRA J. Am. Water Resour. Assoc.* **50**(6), 1401–1424. <https://doi.org/10.1111/jawr.12195> (2014).
42. Rahmani, F., Shen, C., Oliver, S., Lawson, K. & Appling, A. Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. *Hydrol. Processes* **35**(11), e14400. <https://doi.org/10.1002/hyp.14400> (2021).
43. Saha, G., Rahmani, F., Shen, C., Li, L. & Raj, C. A deep learning-based novel approach to generate continuous daily stream nitrate concentration for nitrate data-sparse watersheds. *Sci. Total Environ.* <https://doi.org/10.1016/j.scitotenv.2023.162930> (2023).
44. Souza, J. & Hooke, J. Influence of seasonal vegetation dynamics on hydrological connectivity in tropical drylands. *Hydrol. Processes* **35**(11), e14427. <https://doi.org/10.1002/hyp.14427> (2021).
45. Wang, S., Wang, X. & Ouyang, Z. Effects of land use, climate, topography and soil properties on regional soil organic carbon and total nitrogen in the Upstream Watershed of Miyun Reservoir, North China. *J. Environ. Sci.* **24**(3), 387–395. [https://doi.org/10.1016/S1001-0742\(11\)60789-4](https://doi.org/10.1016/S1001-0742(11)60789-4) (2012).
46. Burns, D. A. *et al.* Monitoring the riverine pulse: Applying high-frequency nitrate data to advance integrative understanding of biogeochemical and hydrological processes. *Wiley Interdiscipl. Rev. Water* **6**(4), e1348. <https://doi.org/10.1002/wat2.1348> (2019).
47. Duncan, J. M., Band, L. E. & Groffman, P. M. Variable nitrate concentration–discharge relationships in a forested watershed. *Hydrol. Processes* **31**(9), 1817–1824. <https://doi.org/10.1002/hyp.11136> (2017).
48. Pardo, L. H., Green, M. B., Bailey, S. W., McGuire, K. J. & McDowell, W. H. Identifying controls on nitrate sources and flowpaths in a forested catchment using a hydrogeological framework. *J. Geophys. Res. Biogeosci.* **127**(2), e2020JG006140. <https://doi.org/10.1029/2020JG006140> (2022).
49. Maggs, J. & Hewett, B. Organic C and nutrients in surface soils from some primary rainforests, derived grasslands and secondary rainforests on the Atherton Tableland in North East Queensland. *Soil Res.* **31**(3), 343–350. <https://doi.org/10.1071/SR9930343> (1993).
50. Zheng, M., Xu, M., Li, D., Deng, Q. & Mo, J. Negative responses of terrestrial nitrogen fixation to nitrogen addition weaken across increased soil organic carbon levels. *Sci. Total Environ.* <https://doi.org/10.1016/j.scitotenv.2023.162965> (2023).
51. Lintern, A. *et al.* The influence of climate on water chemistry states and dynamics in rivers across Australia. *Hydrol. Processes* **35**(12), e14423. <https://doi.org/10.1002/hyp.14423> (2021).
52. Duncan, J. M., Welty, C., Kemper, J. T., Groffman, P. M. & Band, L. E. Dynamics of nitrate concentration–discharge patterns in an urban watershed. *Water Resour. Res.* **53**(8), 7349–7365. <https://doi.org/10.1002/2017WR020500> (2017).
53. Todman, L. C., Bush, A. & Hood, A. S. ‘Small Data’ for big insights in ecology. *Trends Ecol. Evol.* <https://doi.org/10.1016/j.tree.2023.01.015> (2023).
54. Wang, S., Peng, H. & Liang, S. Prediction of estuarine water quality using interpretable machine learning approach. *J. Hydrol.* **605**, 127320. <https://doi.org/10.1016/j.jhydrol.2021.127320> (2022).
55. Karki, N. *et al.* Comparative performance of regionalization methods for model parameterization in ungauged Himalayan watersheds. *J. Hydrol. Reg. Stud.* **47**, 101359. <https://doi.org/10.1016/j.ejrh.2023.101359> (2023).
56. Arora, B. *et al.* Differential CQ analysis: A new approach to inferring lateral transport and hydrologic transients within multiple reaches of a mountainous headwater catchment. *Front. Water* **2**, 24. <https://doi.org/10.3389/frwa.2020.00024> (2020).
57. Lovett, G. M. *et al.* Who needs environmental monitoring?. *Front. Ecol. Environ.* **5**(5), 253–260. [https://doi.org/10.1890/1540-9295\(2007\)5\[253:WNEM\]2.0.CO;2](https://doi.org/10.1890/1540-9295(2007)5[253:WNEM]2.0.CO;2) (2007).
58. Kelleway, J. J. *et al.* Carbon storage in the coastal swamp oak forest wetlands of Australia. *Wetland Carbon Environ. Manag.* <https://doi.org/10.1002/9781119639305.ch18> (2021).
59. Courtois, A., Morel, J. M. & Arias, P. Can neural networks extrapolate? Discussion of a theorem by Pedro Domingos. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A Matemáticas* **117**(2), 79. <https://doi.org/10.1007/s13398-023-01411-z> (2023).
60. Al-Musaylh, M. S., Deo, R. C., Adamowski, J. F. & Li, Y. Short-term electricity demand forecasting using machine learning methods enriched with ground-based climate and ECMWF Reanalysis atmospheric predictors in southeast Queensland, Australia. *Renew. Sustain. Energy Rev.* **113**, 109293. <https://doi.org/10.1016/j.rser.2019.109293> (2019).
61. ABARES. (2016). The Australian Land Use and Management Classification Version 8.
62. Lintern, A. *et al.* What are the key catchment characteristics affecting spatial differences in riverine water quality?. *Water Res. Res.* **54**(10), 7252–7272. <https://doi.org/10.1029/2017WR022172> (2018).
63. The MathWorks Inc., 2020. MATLAB version 2020a [Deep Learning Toolbox]. The MathWorks Inc. [www.mathworks.com](http://www.mathworks.com).
64. Salari, M., Teymouri, E. & Nassaj, Z. Application of an artificial neural network model for estimating of water quality parameters in the Karun river, Iran. *J. Environ. Treat. Techn.* **9**(4), 720–727. [https://doi.org/10.47277/JETT/9\(4\)727](https://doi.org/10.47277/JETT/9(4)727) (2021).
65. Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315–323). JMLR Workshop and Conference Proceedings.
66. Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807–814).

67. Uzair, M., & Jamil, N. (2020, November). Effects of hidden layers on the efficiency of neural networks. In 2020 IEEE 23rd international multitopic conference (INMIC) (pp. 1–6). IEEE.
68. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv* <https://doi.org/10.48550/arXiv.1412.6980> (2014).
69. Deo, R. C. & Şahin, M. Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern Australia. *Atmos. Res.* **161**, 65–81. <https://doi.org/10.1016/j.atmosres.2015.03.018> (2015).
70. Moriasi, D. N. *et al.* Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **50**(3), 885–900. <https://doi.org/10.13031/2013.23153> (2007).
71. Willmott, C. J. Some comments on the evaluation of model performance. *Bull. Am. Meteorol. Soc.* **63**(11), 1309–1313. [https://doi.org/10.1175/1520-0477\(1982\)063%3c1309:SCOTEO%3e2.0.CO;2](https://doi.org/10.1175/1520-0477(1982)063%3c1309:SCOTEO%3e2.0.CO;2) (1982).
72. Fox, D. G. Judging air quality model performance: A summary of the AMS workshop on dispersion model performance, woods hole, Mass., 8–11 September 1980. *Bull. Am. Meteorol. Soc.* **62**(5), 599–609. [https://doi.org/10.1175/1520-0477\(1981\)062%3c0599:JAQMP%3e2.0.CO;2](https://doi.org/10.1175/1520-0477(1981)062%3c0599:JAQMP%3e2.0.CO;2) (1981).
73. Deo, R. C., Tiwari, M. K., Adamowski, J. F. & Quilty, J. M. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. *Stochastic Environ. Res. Risk Assess.* **31**, 1211–1240. <https://doi.org/10.1007/s00477-016-1265-z> (2017).
74. State of Queensland Department of Natural Resources, Mines and Energy (2018) Surface Water Ambient Network (Water Quality) 2018–19, WMP014 version 6, June 2018.

## Acknowledgements

Special thanks to Chris Davey at UniSQ's Advanced Data Analytics Research group for invaluable assistance with coding suited to infrastructure changes in High Performance Computing (HPC) Environments. Also Richard Young for ongoing support and operation of the UniSQ HPC facility. The authors are grateful to the University of Southern Queensland and the Australian Federal Government for the Research Training Program Scholarship and State of Queensland Government supported Queensland Water Quality Modelling Network initiative for top-up funding and motivation. Also, The Bureau of Meteorology for the provision of necessary data, Dr. Urooj Khan for support, and the numerous State of Queensland Government staff who passionately develop, ground truth, and maintain open access spatial datasets to provide unprecedented knowledge of nature.

## Author contributions

C.M.O'S. and A.G. conceptualised the study and methodological approach. C.M.O'S. developed and applied the methodology, processed the data and simulations, analysed and interpreted the results, wrote the original manuscript. A.G. supervised the research. R.D. supervised writing the manuscript. All authors have reviewed, edited and approved submission of this manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-45259-0>.

**Correspondence** and requests for materials should be addressed to C.M.O.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023, corrected publication 2023