EGUSPHERE-2022-617: Author’s response

Developing a Bayesian network model for understanding river catchment resilience under future change scenarios

Kerr J. Adams, Christopher (Kit) A. J. Macleod, Marc J. Metzger, Nicola Melville, Rachel C. Helliwell, Jim Pritchard, Miriam Glendell

Dear Copernicus Publications Production Office,

Please find the point-by-point response to reviewers and list of relevant changes made to the manuscript presented in Table 1. C denotes a reviewer comments.

Reviewer #1:

C1.1) I think a picture of the model should be presented. I understand it can be complex, but I also understand it was presented for the stakeholders in the workshops, so it should be possible to present it also in the paper, or at the minimum in the supplement. It would make it easier for the reader to understand the model.

Thank you for this comment, we accept that a visualisation of the model should be presented. We have included a simplified visualisation as Figure S2 in the Supplementary Material section S3.

Our model contains 417 nodes, 623 arcs and 23 sub-models. Despite not being a spatial model, there are some geographical considerations included to represent the sub-catchment scale and individual wastewater assets, which results in the repetition of nodes, arcs and sub-models. These geographical considerations do make it complex to represent the full model visually, which is why we decided to include a simplified version in the supplementary material. We’ve highlighted where there is repetition in the supporting text box.

We have provided more information on the model structure providing the following text to lines 204-208:

Our model contained 417 nodes, 623 arcs and 23 sub-models. Despite not being a spatial model, there are some geographical considerations included to represent five sub-catchments. Across the five sub-catchments the model included 10 wastewater assets, two public water drinking assets, four land-cover types, four crop types and septic tanks. Dividing the model into sub-catchments resulted in repetition of nodes and arcs.

C1.2) It seems from the supplement that the model was parameterized using deterministic equations. Usually Bayesian Networks are use specifically to model also the uncertainty that is related to the model parameters. Please discuss this and explain your modelling choice.

Thank you for this comment. Where we do have the data available to measure uncertainty, we use prior distributions, which are represented in equations using the β function. The distributions are determined by analysing the available data to identify mean, standard deviations and truncations to create the best possible prior. We also take advantage of a built-in GeNIe function which analyses the data to create a custom prior distribution where there are larger data records to support variables, such as surface water flows. When generating prior distributions isn’t possible, we make use of the diverse coupled future pathways as the best available method for representing uncertainty.

C1.3) The use of simulations to evaluate the results is a bit unclear. We don't usually use simulations as such to evaluate the outputs of a BN, but we aim to compute the total probability distribution over the modelled domain, given the conditional probability distributions and the model structure. This way, we can then reason "backwards" (what is the most probable cause given the consequences), compute the probabilities of outcomes given a number of causes or observations, etc. In the case of discrete models, this can be done analytically, and in the case of continuous models, the distributions are often approximated using simulations, but BNs are not usually simulated as such. When continuous BNs are run/solved, often using Monte Carlo Markov chain computation, the early part of the Markov chain is usually thrown out to make sure that the chain has converged to the true distribution (burn-in). This wasn't mentioned in this paper, and I was left uncertain about the modelling technique. Please explain it more clearly. Also, BNs are supposed to give the best available assessment of the *probabilities* of the events (given the scenarios etc.), so it should not be necessary to refer to "x out of y simulations" whan discussing the results.

Many thanks for highlighting our confusing use of simulations, we will update the manuscript to replace simulation with scenario and samples where appropriate throughout the manuscript as our results are describing and comparing outputs from executing both the current and future (coupled RCPs and SSPs) scenarios.

The modelling technique we use is a hybrid forward sampling algorithm, which is the best available algorithm for hybrid models using the GeNIe software. The algorithm generates 10,000 samples and when stakeholders enquired what was meant by, for example, a 51% probability of a variable being resilient.

We have added details of the forward sampling algorithm in the methods section lines 263-269 of the manuscript explaining the following:

The model was updated using the default GeNIe software hybrid forward sampling algorithm. The algorithm computes 10,000 samples from the prior probability distributions of parentless nodes, which it then used to generate samples in child nodes of the prior parent node distribution(s), generating probability distributions. Summary statistics (mean, standard deviation, minimum and maximum) were derived from the probability distributions for each node, which were compared for different current and future pathway scenarios.

4) Maybe go further back to the roots (such as Perl 1986) when explaining what BNs are in the introduction.

Reference is now made to the work of Pearl (1986) describing BNs as directed acyclic graphs and conditional probability quantification, line 49.

Reviewer #2

C2.1) The authors use the term resilience but do not provide a clear definition on what that term is supposed to capture in their work.

Thank you for this comment. We agree that a definition of resilience is lacking in the manuscript.

We will include the following text to define resilience and make it clear that we are capturing socio-ecological resilience in the manuscript from lines 13-19:

Resilience was first introduced by (Holling, 1973) as the ability of ecological systems to absorb disturbances and retain their functions in the face of change. Adger (2000), later defined social resilience as the ability of groups and communities to cope with social, political and environmental change. The crossover between social and ecological theories led to the theory of socio-ecological system resilience (Cretney, 2014, Folke, 2006). Decision-makers must be able to understand how a system shifts from one state to another (Renaud et al., 2010) to inform resilient water management and allow freshwater systems to bounce back and adapt to variability, uncertainty and transformation (Brown, 2015).

C2.2) The authors need to provide references to the One Planet Choices work given its tight coupling with their work.

Thank you for this comment. We have provided the following reference to line 104 of the manuscript:

"SEPA: Sustainable Growth Agreement Scottish Water and Scottish Environment Protection Agency Progress Update February. 2020. Available Online: https://www.sepa.org.uk/media/496202/scottish-water-sga-update.pdfhttps://www.sepa.org.uk/media/496202/scottish-water-sga-update.pdf.

We have also included a footnote next to the first reference of One Planet Choices in line 96 with the following link to a video describing the method: https://vimeo.com/804313679/1139d31b45"

C2.3) How did you ensure that when you were eliciting the model from the stakeholders that you did not end up with cyclical pathways?

Thank you for highlighting these questions. We agree that methods used to prevent cyclical pathways need to be made clearer in the manuscript.

The model section headings previously presented in lines 156-162 were used to ensure that the purpose of the BN model was clear to the different stakeholder groups. In addition, they were also used to ensure that cause-and-effect relationships were linear, which is evident in Figure 3 (line 352) in the previous manuscript and our response to Referee #1 comment C1.1.

The model section headings were used during stage two of the five-stage participatory approach described in Figure 3 line 198 of the manuscript to keep the model linear. We provide a visual representation of the headings in Figure 2 line 166 and updated lines 163-165 of the original manuscript to the following:

Model section headings (Figure 2) were agreed with the project team at the outset to clarify the modelling purpose with different stakeholder groups and ensure that the elicited cause-and-effect relationships were linear.

C 2.4) The paper needs to discuss the choice of the spatio-temporal scales in more details. The data is reported using daily units; yet the model is predicting a yearly average in 2050. The authors work with subwatersheds, some of which are nested. How was the nesting accounting for?

Thank you for these comments and questions. Model outputs provide the average daily unit loads/concentration values for 2050, rather than the average yearly load/concentration values. We use the equation-based BN model to investigate how different climate, population and land cover scenarios to 2050 would change daily unit loads/concentrations and present the outputs as the average daily load/concentration to 2050. For example, the 'annual' scenario takes account of the predicted annual precipitation rate anomaly, therefore in the 'annual' scenario we investigate how the average daily rate might change given the average annual predicted change in precipitation.

Sub-watersheds were modelled using sub-models to create a semi-distributed model. The sub-models were nested to account for the routing of water and contaminants from upstream sub-watersheds to downstream sub-watersheds.

The spatial and temporal scales are limited in the model meaning many of the variables are static in time and space. In this research, we aimed to address gaps in the application of equation-based hybrid BN structures in environmental risk assessment. Given the findings from the research, we would like to understand ways of developing a spatial and dynamic BN model in the future, however, this was not the focus of the current research described in the manuscript.

We specifically address the need to discuss the choice of spatio-temporal scale in more detail in the paper in our response to comment 2.5 of the review.

C2.5) Many of the variables in the model are static in time and space. For example all subwatersheds had the same projected change in precipitation.

Thank you for the comments and questions. It is correct that many of the variables in the model are static in time and space. Using the UKCP18 precipitation change anomaly data for a 25 km grid square meant that the entire catchment was within the same spatial grid and therefore had the same projected change in precipitation. If we are to investigate the use of a spatial BN model in the future, we could utilise the UKCP18 2.2 km grid square product. We believe this should be considered in the manuscript and have included the following text from line 674 of the manuscript:

For example, in this study, we considered future precipitation change anomalies using the UKCP18 25 km grid square data which is limited compared to the possible use of UKCP18 2.2 km grid square precipitation change anomaly data.

All subwatersheds had the same P application, etc.. The authors need to have a table that highlights what were the variables that varied by time and by watershed and what was assumed to not vary over space and/or time.

The P application does differ in the different watersheds based on diffuse Phosphorus loads in kg/day, derived from the ADAS PSYCHIC (Phosphorus and Sediment Yield Characterisation In Catchments) model by Davison et al. (2008) that uses land use data (AgCensus) to estimate loads derived from Arable and Livestock land cover in 1km grids used in a Source Apportionment GIS model. Given we knew the land cover of arable land in each sub-catchment we could calculate the kg/ha/day load, which is presented as γ value in supplementary table S3.

We accept there needs to be greater clarity on the variable spatial and temporal scales. To achieve this Table S3 of the Supplementary Material has been updated to describe if a variable is static in space and time.

How did you define the spatial and temporal scales? Did your stakeholders agree on the adopted spatio-temporal resolution?

Stakeholders co-developed the conceptual model structure, providing details of the different sub-catchments and assets in the catchment. Stakeholders then provided the data and metric information to support each of the variables included in the model structure to inform the spatio-temporal resolution.

We have added the following detail in lines 192-193 of the manuscript to make this point clearer:

Data, metric and catchment specific information provided by stakeholders for each model variable informed the spatio-temporal resolution of the model.

The authors use the Precipitation rate anomaly as the major CC metric. This might be good for point sources; for no-point sources the annual % change is not the best option since P application is seasonal. Please explain the decision to move to annual rather than a seasonal scale.

We agree that exploring using annual changes is limiting and that seasonal considerations would be best for no-point sources, however, to consider a seasonal scale would require data on the seasonal differences in Phosphorus applications in the catchment, which were not available to us. We only had access to static diffuse P loads in kg/day, from the ADAS PSYCHIC model. If seasonal diffuse P data was available, it would be possible to consider seasonal differences, as the UK Climate Projection data does provide seasonal precipitation rate anomalies, which are considered in extreme scenarios.

Despite the data limitation, we would like to consider changes in seasonal Phosphorus applications while also responding to referee questions provided in comment C2.12 regarding the consideration of the sensitivity of the model to input parameters. We have included model outputs for a 20% increase/decrease in P application rates in each waterbody sub-catchment. The outputs of the sensitivity analysis are presented in Table 3 of the manuscript.

We have combined a full response in comment C2.12.

C2.6) The authors need to explain more their adopted methodology that they used to model LULC change over time. Was the change over time assumed to be linear? Did they track the change as a function of what the original LULC class was?

Thank you for this important comment. LULC change was assumed to be linear until 2050, tracking the change as a function of the baseline LULC class cover. We used a story and simulation approach to consider how the land cover would change across different regional shared socioeconomic pathways (SSPs) for the UK compared to the current land cover as the baseline. Having reviewed the information in the manuscript, we believe greater information on how SSP narrative storylines were used to support the story and simulation process should be provided with land cover change values.

To better represent the land cover change in each waterbody sub-catchment under the different scenarios we have removed Table S6 and Figure S4 of the original manuscript supplementary material and replaced them with Figures S4-S8 in Supplementary Material section S4.

We update the manuscript to for details on where to find the supplementary information in lines 252-255:

The percentage cover was converted to hectares (Ha) for each land cover type in each of the waterbody sub-catchments (S4, Figure S4-S8). Projected land cover change values in comparison to 2019 land cover for the entire catchment are provided in S4 Figure S9. Section S4 includes a detailed description of how land cover values were derived.

We include the following text in S4 of the Supplementary Material:

We used the Shared Socioeconomic pathway scenarios developed by Pedde et al (2021) to use the trends of how land cover may change in the future using UK-SSP1 as the basis for the Green Road Scenario, UK-SSP2 for the Business as Usual Scenario and UK-SSP5 as the basis for the Fossil Fuelled Development Scenario.

In the Green Road scenario, there is a greater emphasis on protecting environmental areas, therefore an increase in woodland and wild grasslands is evident in the scenario. The Green Road Scenario considers a switch to a more vegetarian-based diet, resulting in the reduction of pasture land for meat production. In contrast, the Fossil-Fuelled Development scenario includes less consideration for environmental protection and maximises the amount of land in traditional economic-based land covers, such as pasture, to support a meat-based diet and conifer plantations. Using local interpretations from stakeholders, it was clear that arable farming was a key source of income in the catchment and an area of Scotland highly desirable for arable farming, therefore, the arable land cover was increased across all scenarios. For urban land cover, all scenarios considered an increase in population in Cupar, which is the largest town in the catchment, and the neighbouring Foodieash and Dairsie. Less urbanisation is considered in the Green Road scenario as living in rural areas is considered more desirable in comparison to the Fossil Fuelled Development scenario, which sees the greatest increase in urbanisation as the UK-SSP5 trends predict an increase in movement to eastern Scotland.

For the Business as Usual scenario, land cover trends from 1990 were used to determine the changes in land cover. Arable cover has been the predominant type since the 1990s and has been gradually increasing. Pasture land has the second largest coverage, but has been gradually declining. Trends since 2010 were used to inform the Business as Usual trajectory to 2050 and historic values from the 1990s were used to consider the upper limits of the different land cover types through time.

Understanding the historic land cover type helped inform the story and simulation approach to inform the boundaries of how the land cover could change in the future under each scenario. The total hectares in each sub-catchment per land cover type were calculated for current conditions (2019) (Morton et al., 2020). The percentage of each land cover type in each sub-catchment was then calculated and altered based on the different scenario narratives, local stakeholder knowledge and historical boundaries, before being converted back to hectares. The different land cover hectares across the different scenarios are presented in Figures S4-8. There are only subtle differences, particularly in arable land cover in the different scenarios, mainly due to the arable land cover nearly being maximised in the catchment currently.

C2.7) The authors adopt a 10,999 monte carlo simulations, Is that enough given the complicated model? How did they know that this value is good enough?

Thank you for this question. We tested running the model using 1,000, 10,000 and 100,000 simulations and check for differences in model output values for reactive phosphorus concentrations (µg/l) in sub-catchment 6200 (the catchment outlet). As there was no difference in model output values between these scenarios, we decided to use the standard 10,000 simulations applied in GeNIe software.

C2.8) Page 11 line 254: Thee authors adopted the equal interval discretization. Why that and not equal frequency? Did you run a sensitivity analysis to determine the impact of the type of discretization on model results?

Thanks for this consideration. We use equal intervals to represent the final output nodes on a discrete scale to support stakeholders in intuitively understanding discretised outputs. However, the model itself is run as a hybrid model, not a fully discretised model. The intervals are informed using an indexing method and regulatory standards where available.

C2.9) Line 266: "sum of the 'IF' statement was used to determine their overall state." This assumes all nodes are all equal. How can you defend that? For example, if you have 3 nodes and each have a score of one is that the same as having one node with a score of 3 and 2 nodes with a score of zero. Is that accurate? How did your stakeholders feel about these assumptions?

Thank you for this comment and the example provided. The IF statement indexing method is used for variables such as natural capital which don't have a common metric for measurement. The first point to raise is that finding common ground across multiple sectors to agree on which capital and their resources should hold greater value is complex. An environmental regulator may value water quality to be the most important variable in the system, while a food producer may value soil quality or financial margins as holding greater value. To reduce complexity, we assumed that all nodes, particularly capitals and their capital resources, are equal in value.

We developed the IF statement indexing method and presented it to the stakeholder 'project team' who accepted the assumption of the index scoring. The 'project team' explained that a 'one out, all out' approach, described by Carvalho et al., (2019), is used when recording ecological status as part of the Water Framework Directive reporting. This method does have the disadvantage in that two goods don't equal one bad, which can lead to regulators being set up to fail, however, it was highlighted that taking a precautionary approach prevents the masking of undesirable outcomes when averaging out scores and provides an easy and transparent way of measuring overall variable states.

We have updated lines 288-293 of the manuscript to reflect on this justification:

As multiple parent nodes were associated with capital and capital resource variables, the sum of the 'IF' statement was used to determine their overall state. The 'IF' statement indexing method follows the 'one out, all out' approach applied to the evaluation of Good Ecological Status in the EU Water Framework Directive, as described in Carvalho et al., (2019). The 'one out all out' approach adopts the precautionary principle to prevent masking of undesirable outcomes when averaging scores and provides an easy and transparent way of measuring overall variable states.

C2.10) Line 269: There is no Appendix E:

Thank you for highlighting this, our apologies for the error. We have updated the manuscript to reference S5 of the supplementary material instead of Appendix E.

C2.11) Lines 278-281: Expand on how the model valuation was done. Did you compared the mean? Did you averaged concentration values over time? If so over which period? Did you see the sd? Was this done for all subwatersheds? What metric did you use to evaluate?

Thanks for this comment, we accept the information regarding the goodness of fit evaluation is unnecessarily split between sections 2.2.4 (methods) & 3.5 (results) of the manuscript, which relates to points raised in comment C2.15.

Using 52 reactive phosphorus concentrations (µg/l) observations at the 6200 sub-catchment outlet collected between 2017-2019, we fitted a histogram using the custom function tool in GeNIe to create an 'observed phosphorus concentration (µg/l) 6200' variable in the model, which was both parentless and childless. The goodness of fit method was done only in 6200 as this is the catchment outlet, therefore all other sub-catchment will drain to this point, making it a suitable point to evaluate.

When running the 'current' model scenario, we compared the median, standard deviation and discretised class probabilities – informed by the WFD classification boundaries for the sub-catchment – for both the modelled reactive phosphorus concentrations and observed reactive phosphorus variables to evaluate the goodness of fit. In a relevant study since the submission of our preprint, Glendell et al., (2022) applied a % Bias measure of model performance where the departure of +/-50% from observations was considered behavioural, we have therefore updated our evaluation to include this method as an additional metric to evaluate the model.

Addressing comments C2.11 and C2.15 we have updated the manuscript from lines 307-326 with the following:

Model performance was evaluated using a goodness of fit method (Aguilera et al., 2011) using 52 bi-monthly observed RP concentrations in micrograms per litre (µg/l) collected in sub-catchment 6200 collected between 2017-2019, (Scottish Water, 2020). We fitted a histogram using the custom function tool in GeNIe to create an 'observed phosphorus concentration (µg/l) 6200' variable, which was both parentless and childless. We evaluated sub-catchment 6200 as this is the catchment outlet for all sub-catchments. Computing the 'current' model scenario, we compared the median, standard deviation and discretised class probabilities – informed by the WFD classification boundaries for the sub-catchment – for both the modelled RP concentrations and observed RP variables to evaluate model goodness of fit.

A % Bias method (Eq.1) applied by Glendell et al., (2022), with a departure of +/−50% from observations considered behavioural, was used to further evaluate model performance:

(Eq. 1) *%Bias= (X_sim-X_obs)/X_obs*

Where X_sim is the modelled RP concentration (µg/l) and X_obs is the observed RP concentration (µg/l).

A one-at-a-time parameter sensitivity analysis was conducted to determine which input variables contributed the greatest variability to model outputs (Wohler et al., 2020, Hamby, 1994). We used the target variable RP concentrations (µg/l) at the 6200 catchment outlet to determine the sensitivity of the model to diffuse pollution phosphorus loads and point source wastewater phosphorus loads. The sensitivity analysis compared the median RP concentration (µg/l) for the current scenario against the +/- 20% difference for diffuse arable, pasture and septic tank P sources, and wastewater P sources while holding other input values constant.

Results reported in the manuscript from lines 483-497 have been updated to the following:

We evaluated the model performance by comparing the modelled current RP concentrations (µg/l) with a simulation of observed RP concentrations (µg/l) at the catchment outlet in waterbody sub-catchment 6200 (Table 1). The model underestimated the median RP concentration (157.63 µg/l) at the catchment outlet compared to the observed simulated median RP concentration (168.82 µg/l). A greater standard deviation was observed in the model simulation (361.7 µg/l) compared to the observed simulation (109.3 µg/l).

Based on the discrete output (Figure 10), the model underestimated the RP concentration compared to the observed simulation. The most probable state for RP concentrations in the observed simulation was moderate risk (44% probability) - or poor WFD status - compared to the modelled scenario which estimated low-risk - or moderate ecological status – (41% probability). The modelled RP concentrations were more widely distributed, which is evident in a 2% probability of high-risk - or bad ecological status - compared with 0% in the observed simulation.

When evaluating the goodness of fit using the % bias correction (Table 2) 43% of observations were within the +/- 50% behavioural threshold, 31% of simulated values were above the 50% acceptable threshold, and 26% were below the 50% acceptable threshold.

C2.12) The authors are encouraged to do a sensitivity analysis (sensitivity to findings and sensitivity to parameters)

Thank you, we appreciate the comment. We didn't discretise the model and perform a sensitivity to findings analysis to avoid loss of information and imprecision discussed in lines 598-600.

We do agree that a sensitivity to parameters investigation would improve the manuscript. The different Representative Concentration Pathways and Shared Socioeconomic Pathways provide a measure of sensitivity to changes in input variables Climate, Population and Land Cover, however, there is limited sensitivity testing to changes in diffuse and point source pollution concentrations, as highlighted in response to comment C2.5.

We have updated the manuscript to include details of the parameter sensitivity analysis of diffuse and point source variables in section 2.2.4 from lines 320-326:

A one-at-a-time parameter sensitivity analysis was conducted to determine which input variables contributed the greatest variability to model outputs (Wohler et al., 2020, Hamby, 1994). We used the target variable RP concentrations (µg/l) at the 6200 catchment outlet to determine the sensitivity of the model to diffuse pollution phosphorus loads and point source wastewater phosphorus loads. The sensitivity analysis compared the median RP concentration (µg/l) for the current scenario against the +/- 20% difference for diffuse arable, pasture and septic tank P sources, and wastewater P sources while holding other input values constant.

We then present findings from the sensitivity analysis in section 3.5 of the results in lines 498-503 onwards with the following text and additional Table 3

The results of the parameter sensitivity analysis are presented in Table 3. Changes in point source RP loads have a greater influence on RP concentrations (µg/l) compared to diffuse sources in sub-catchment 6200 in the current scenario. A 20% increase in point source loads resulted in an 8.4% increase in RP concentrations, while a 20% reduction resulted in an 8.1% reduction in concentrations. Of the diffuse sources, arable sources had the greatest influence on RP concentration with a 20% increase yielding a 4.9% increase in concentration, while a 20% reduction resulted in a 6.5% reduction in concentrations.

C2.13) Figure 6: Discuss why the change in 6202, 6205, 6206 is so high and different from the rest under scenario (d).

Thank you for this comment. The change in reactive phosphorus concentrations in sub-catchments 6202, 6205 and 6206 in scenario (d) (now Figure 7) is because scenario (d) assumes an extreme high rainfall scenario which results in both increased diffuse source run-off and increase effluent loads as there is a greater probability of effluent discharge and spills. We believe the change is higher in the three sub-catchments compared to sub-catchments 6200 and 6201 because we can’t consider the influence of river flow volume and its diluting influence on reactive phosphorus concentrations. River flow volume data was not available for sub-catchments 6202, 6205 and 6206, only sub-catchments 6200 and 6201.

Another reason behind the higher change was due to results in the previous version of the manuscript reporting mean concentration values, which are influenced by outliers. The Fossil Fuelled Development Extreme High Precipitation scenario has the greatest standard deviation, as is evident in the graphs in our response to comment C2.14.

We have since updated the manuscript to present median statistics instead of mean statistics (more information in response to comment C2.16). The figure has been updated in the manuscript as Figure 7.

In addition, we have added the following text to the manuscript in lines 621-655:

Investigations of future scenarios highlighted that in the future BAU scenario (Figure 7, Pane b) median RP concentrations (µg/l) increased compared to current conditions in sub-catchments 6200, 6201 and 6205 and decreased in sub-catchments 6202 and 6206. Figure 8 for sub-catchment 6200 (and Figures S8-12) show increases in total RP loads (kg/day) in sub-catchments 6200, 6201 and 6205, while the total RP loads in sub-catchment 6202 and 6206 decreased, particularly for wastewater sources. The changes in total RP can be seen in the source apportionment between wastewater and diffuse sources, as well as the trends in climate, population and land cover change. Wastewater sources increase in sub-catchments where the population is projected to increase, while diffuse sources are expected to increase in all sub-catchments.

In the Green Road and Fossil-Fuelled Development Extreme Precipitation scenarios, the influence of precipitation change and catchment processes are evident. Total RP loads (kg/day) are reduced in all sub-catchments in the GR ExLP scenario due to reductions in diffuse run-off. The lower likelihood of wastewater spills contributing untreated effluent to wastewater source loads are also reduced in the GR ExLP scenario. RP concentrations (µg/l) were greater in the GR ExLP scenario compared to the current scenarios in sub-catchments 6200 and 6201, despite the reductions in total RP loads in both sub-catchments (Figure 8 and Figures S8-12). We believe these concentration increases are due to the reduction in river flow volumes in the extreme low precipitation rate scenario, meaning regulating diluting functions are absent and RP concentrations increase. We are unable to investigate the influence of flows in the sub-catchments where RP concentrations decreased compared to current conditions (6202, 6205 and 6206) as observed river flow volume data were not available for all sub-catchments (see SM Table 2 for more information on how surface water quality is measured absence of river flow volume data).

In the FFD ExHP scenario, increases in RP concentrations (µg/l) compared to current conditions are evident in all sub-catchment waterbodies, which is attributed to increases in total RP loads (kg/day). Increased precipitation rates increase diffuse run-off, wastewater effluent flows and the likelihood of effluent spills. For sub-catchments 6200 and 6201, despite increases in river flow volumes from increased precipitation, RP source loads into the waterbodies was greater than the dilution capacity.

Despite 46% of the % bias observations falling within the +/- 50% acceptable model performance (Table 2), results from the goodness of fit evaluation demonstrate that the model underestimated current median RP concentrations (µg/l) at the catchment outlet in sub-catchment 6200 and the probable risk class. Simulated concentrations were more widely distributed, as compared to the observed data, as is evident in the 2% of observations within a high-risk state for simulated concentrations, compared to 0% for observed concentrations. A wider distribution in simulated RP values using a hybrid BN model was also found by Glendell et al., (2022). We concur with their considerations that both the quality and the low temporal resolutions of observed data may be responsible for this discrepancy.

C2.14) Bar charts 7 and 8 need to show the uncertainty bounds. Also provide similar charts for the rest of the subwatersheds in the SM

Thank you for this comment, we have included the Bayesian credible interval (Q5 and Q95) uncertainty bounds in both figures (now Figures 8 and 9 in the updated manuscript) and added the results for the remaining sub-catchments in Figures S10-13 section S6 of the Supplementary Material.

C2.15) Line 447: Explain why the evaluation was done in sub-catchment 6200. Also state the period over which the data was collected and used for that sub-catchment.

Thank you for this comment, we have acknowledged and addressed this in our response to comment C2.11 regarding sub-catchment 6200 being the catchment outlet.

C2.16) Figure 9: Did you check and see if these differences are coming from the point or non-point sources. You have the point source data for all WWTPs so you can estimate the relative contribution of each.

Thank you for this consideration. We have now considered this in our sensitivity to input parameters analysis described in response to comment C2.12 however, we believe that this comparison wouldn't provide a complete justification as to why the model underestimates the probability of being within a moderate risk class.

We have updated the manuscript to report results as median values, rather than mean values as there are outliers in the modelled outputs, which was influencing the overestimation values in Table 1 compared to the underestimation in the discretised outputs in Figure 9 of the original manuscript.

Further, we accept that a discussion of the underestimation and wider distribution evident for the modelled reactive phosphorus concentrations (µg/l) is required in the manuscript. We have included the following discussion on the potential reasoning behind the underestimation of the modelled reactive phosphorus concentrations and the observed concentrations from line 649-656 of the manuscript:

Despite 46% of the % bias observations falling within the +/- 50% acceptable model performance (Table 2), results from the goodness of fit evaluation demonstrate that the model underestimated current median RP concentrations (µg/l) at the catchment outlet in sub-catchment 6200 and the probable risk class. Simulated concentrations were more widely distributed, as compared to the observed data, as is evident in the 2% of observations within a high-risk state for simulated concentrations, compared to 0% for observed concentrations. A wider distribution in simulated RP values using a hybrid BN model was also found by Glendell et al., (2022). We concur with their considerations that both the quality and the low temporal resolutions of observed data may be responsible for this discrepancy.

C2.17) It is not clear why the authors adopted a continuous BN to only discretize its output. An explanation on the reasons for that is needed.

Thank you for this comment. Firstly, we avoided the discretisation of continuous variables within the model to avoid loss of information. However, from the stakeholder's perspective, understanding the probability of model outputs falling into agreed risk classes helped the understanding and the communication of the results. Therefore, we discretised terminal continuous nodes to provide a dual representation of outputs to support effective communication of findings, as described in lines 542-549 of the original manuscript.

To make the reasoning clearer we will update the manuscript as follows in lines 279-271:

"We presented a dual representation of continuous nodes using a discretised child node to support the communication of the results using both summary statistics (median and standard deviation) available in continuous outputs and the probability of model outputs falling into agreed risk classes available in discrete variables."

Providing a simplified visualisation of the BN model in response to Referee 1 comment C1.1 supports our additional comments.

C2.18) Discuss why the GeNIe platform was used as compared to others?

Thanks for this consideration. We used the GeNIe software because it was free for academic use and it allows the development of hybrid models without discretising the output. Further, members of the author team have experience in using GeNIe, for example in Stewart et al, 2021 and Glendell et al 2022, meaning an in-depth understanding of the software was already available to the research team.

C2.19) Table S2 in the SM is very important; yet it hard to follow. It also needs English editing.

Thank you for this comment, we accept that there is a lot of detail in Table S2 which makes it difficult to follow. We included as much detail about the nodes within the model as possible to provide transparency, however, the depth of detail may have been counterintuitive.

To address this, we updated Table S2 by combining node name and identifier columns, removing state and discretisation columns as they are presented in Table S3 and simplifying the supporting information column to only include key information.

C2.20) Are the gammas in Table S2 fixed or do they have distributions around them? How were these determined?

Thank you for these questions. The gamma values are fixed values, which are available in Table S3. The values are determined using a mixture of methods which are explained in the supporting information column of Table S2. In our response to comment C2.5, we have tried to make the justification for gamma values clearer when updating Table S3.

C2.21) The authors limit their continuous distributions to normal and truncated normal. Why?

Thanks for this question. We mainly use normal and truncated normal distributions because there is limited data to use/fit different distribution types. For example, for wastewater phosphorus concentrations we rely on mean and standard deviation values, rather than multiple observations or data points. Where we do have sufficient data points/observations, we use the 'custom' GeNIe function, which fits a customised histogram distribution to the data available, which is the preferred method but not always possible with the reliance on summary statistics. Hence, in that sense, our model acts as a meta-model integrating output from proprietary models as well as regulatory data.

We added further clarification to the manuscript in lines 257-263:

Where supporting continuous data was available, we fitted truncated normal prior distributions by calculating the mean and standard deviation from available values. Truncated normal distributions were fitted to avoid negative values, where appropriate. Secondly, where longer data records were available, we used a built in GeNIe function to fit a custom prior distribution (histogram) to time-series data. Where available data was limited to a single deterministic value and statistical moments could not be calculated, we applied scenario modelling using the diverse coupled future pathways as a best available method for representing uncertainty.

C2.22) Table S3: Make sure you repeat header on all pages.

Thank you for highlighting this, we have repeated the header for Tables S3 and S2.

List of relevant changes:

Reviewer comment	Change type	Changes made to manuscript	Location in manuscript
C1.1	Additional Supplementary Material	Add simplified visualisation of the Bayesian Network model.	Figure S2 in the Supplementary Material section S3

C1.1	Additional text	*Our model contained 417 nodes, 623 arcs and 23 sub-models. Despite not being a spatial model, there are some geographical considerations included to represent five sub-catchments. Across the five sub-catchments the model included 10 wastewater assets, two public water drinking assets, four land-cover types, four crop types and septic tanks. Dividing the model into sub-catchments resulted in repetition of nodes and arcs.*	Lines 204-208
C1.3	Additional text	*The model was updated using the default GeNIe software hybrid forward sampling algorithm. The algorithm computes 10,000 samples from the prior probability distributions of parentless nodes, which it then used to generate samples in child nodes of the prior parent node distribution(s), generating probability distributions. Summary statistics (mean, standard deviation, minimum and maximum) were derived from the probability distributions for each node, which were compared for different current and future pathway scenarios.*	Lines 263-269
C1.4	Additional reference	Pearl (1986)	Line 49 and reference list.
C2.1	Additional text	*Resilience was first introduced by (Holling, 1973) as the ability of ecological systems to absorb disturbances and retain their functions in the face of change. Adger (2000), later defined social resilience as the ability of groups and communities to cope with social, political and environmental change. The crossover between social and ecological theories led to the theory of socio-ecological system resilience (Cretney, 2014, Folke, 2006). Decision-makers must be able to understand how a system shifts from one state to another (Renaud et al., 2010) to inform resilient water management and allow freshwater systems to bounce back and adapt to variability, uncertainty and transformation (Brown, 2015).*	Lines 13-19
C2.2	Additional reference and footnote	(SEPA, 2020) and footnote 1: A visual description of the One Planet Choices approach can be found by following this link.	Lines 103-104
C2.3	Additional text and figure	*Model section headings (Figure 2) were agreed with the project team at the*	Lines 163-166 Figure 2 on line 166

		outset to clarify the modelling purpose with different stakeholder groups and ensure that the elicited cause-and-effect relationships were linear.	
C2.5	Additional text	*For example, in this study, we considered future precipitation change anomalies using the UKCP18 25 km grid square data which is limited compared to the possible use of UKCP18 2.2 km grid square precipitation change anomaly data.*	Line 674-676
C2.5	Updated Supplementary Material	Updated Table S3 of the supplementary material to note if each node in the Bayesian Network model included spatial or static data.	Table S3 in Supplement in the Supplementary Material section S3.
C2.5	Additional text	*Data, metric and catchment specific information provided by stakeholders for each model variable informed the spatio-temporal resolution of the model.*	Lines 192-193
C2.6	Edited text	*The percentage cover was converted to hectares (Ha) for each land cover type in each of the waterbody sub-catchments (S4, Figure S4-S8). Projected land cover change values in comparison to 2019 land cover for the entire catchment are provided in S4 Figure S9. Section S4 includes a detailed description of how land cover values were derived.*	Lines 252-255
C2.6	Additional Supplementary Material Figures	Added figures of land cover type (Ha) differences between scenarios in each waterbody sub-catchment.	Figures S4-8 in Supplementary Material section S4.
C2.6	Additional Supplementary Material Text	*We used the Shared Socioeconomic pathway scenarios developed by Pedde et al (2021) to use the trends of how land cover may change in the future using UK-SSP1 as the basis for the Green Road Scenario, UK-SSP2 for the Business as Usual Scenario and UK-SSP5 as the basis for the Fossil Fuelled Development Scenario.* *In the Green Road scenario, there is a greater emphasis on protecting environmental areas, therefore an increase in woodland and wild grasslands is evident in the scenario. The Green Road Scenario considers a switch to a more vegetarian-based diet, resulting in the reduction of pasture land for meat production. In contrast, the Fossil-Fuelled Development scenario includes less*	Supplementary Material section S4.

		consideration for environmental protection and maximises the amount of land in traditional economic-based land covers, such as pasture, to support a meat-based diet and conifer plantations. Using local interpretations from stakeholders, it was clear that arable farming was a key source of income in the catchment and an area of Scotland highly desirable for arable farming, therefore, the arable land cover was increased across all scenarios. For urban land cover, all scenarios considered an increase in population in Cupar, which is the largest town in the catchment, and the neighbouring Foodieash and Dairsie. Less urbanisation is considered in the Green Road scenario as living in rural areas is considered more desirable in comparison to the Fossil Fuelled Development scenario, which sees the greatest increase in urbanisation as the UK-SSP5 trends predict an increase in movement to eastern Scotland. *For the Business as Usual scenario, land cover trends from 1990 were used to determine the changes in land cover. Arable cover has been the predominant type since the 1990s and has been gradually increasing. Pasture land has the second largest coverage, but has been gradually declining. Trends since 2010 were used to inform the Business as Usual trajectory to 2050 and historic values from the 1990s were used to consider the upper limits of the different land cover types through time.* *Understanding the historic land cover type helped inform the story and simulation approach to inform the boundaries of how the land cover could change in the future under each scenario. The total hectares in each sub-catchment per land cover type were calculated for current conditions (2019) (Morton et al., 2020). The percentage of each land cover type in each sub-catchment was then calculated and altered based on the different scenario narratives, local stakeholder knowledge and historical boundaries, before being converted back*	

		to hectares. The different land cover hectares across the different scenarios are presented in Figures S4-8. There are only subtle differences, particularly in arable land cover in the different scenarios, mainly due to the arable land cover nearly being maximised in the catchment currently.	
C2.9	Additional text	*As multiple parent nodes were associated with capital and capital resource variables, the sum of the 'IF' statement was used to determine their overall state. The 'IF' statement indexing method follows the 'one out, all out' approach applied to the evaluation of Good Ecological Status in the EU Water Framework Directive, as described in Carvalho et al., (2019). The 'one out all out' approach adopts the precautionary principle to prevent masking of undesirable outcomes when averaging scores and provides an easy and transparent way of measuring overall variable states.*	Lines 288-293
C2.10	Edited text	*A detailed example of the IF statement indexing method is provided in S5 of the supplementary material.*	Lines 296-296
C2.11	Additional text	*Model performance was evaluated using a goodness of fit method (Aguilera et al., 2011) using 52 bi-monthly observed RP concentrations in micrograms per litre (µg/l) collected in sub-catchment 6200 collected between 2017-2019, (Scottish Water, 2020). We fitted a histogram using the custom function tool in GeNIe to create an 'observed phosphorus concentration (µg/l) 6200' variable, which was both parentless and childless. We evaluated sub-catchment 6200 as this is the catchment outlet for all sub-catchments. Computing the 'current' model scenario, we compared the median, standard deviation and discretised class probabilities – informed by the WFD classification boundaries for the sub-catchment – for both the modelled RP concentrations and observed RP variables to evaluate model goodness of fit.*	Lines 307-326

		A % Bias method (Eq.1) applied by Glendell et al., (2022), with a departure of +/−50% from observations considered behavioural, was used to further evaluate model performance: (Eq. 1) %Bias= (X_sim-X_obs)/X_obs Where X_sim is the modelled RP concentration (µg/l) and X_obs is the observed RP concentration (µg/l). A one-at-a-time parameter sensitivity analysis was conducted to determine which input variables contributed the greatest variability to model outputs (Wohler et al., 2020, Hamby, 1994). We used the target variable RP concentrations (µg/l) at the 6200 catchment outlet to determine the sensitivity of the model to diffuse pollution phosphorus loads and point source wastewater phosphorus loads. The sensitivity analysis compared the median RP concentration (µg/l) for the current scenario against the +/- 20% difference for diffuse arable, pasture and septic tank P sources, and wastewater P sources while holding other input values constant.	
C2.11	Additional text and table	We evaluated the model performance by comparing the modelled current RP concentrations (µg/l) with a simulation of observed RP concentrations (µg/l) at the catchment outlet in waterbody sub-catchment 6200 (Table 1). The model underestimated the median RP concentration (157.63 µg/l) at the catchment outlet compared to the observed simulated median RP concentration (168.82 µg/l). A greater standard deviation was observed in the model simulation (361.7 µg/l) compared to the observed simulation (109.3 µg/l). Based on the discrete output (Figure 10), the model underestimated the RP concentration compared to the observed simulation. The most probable state for RP concentrations in the observed simulation was moderate risk (44%	Lines 483-497 and Table 2 line 520

		probability) - or poor WFD status - compared to the modelled scenario which estimated low-risk - or moderate ecological status – (41% probability). The modelled RP concentrations were more widely distributed, which is evident in a 2% probability of high-risk - or bad ecological status - compared with 0% in the observed simulation. *When evaluating the goodness of fit using the % bias correction (Table 2) 43% of observations were within the +/- 50% behavioural threshold, 31% of simulated values were above the 50% acceptable threshold, and 26% were below the 50% acceptable threshold.*	
C2.12	Additional text	*A one-at-a-time parameter sensitivity analysis was conducted to determine which input variables contributed the greatest variability to model outputs (Wohler et al., 2020, Hamby, 1994). We used the target variable RP concentrations (µg/l) at the 6200 catchment outlet to determine the sensitivity of the model to diffuse pollution phosphorus loads and point source wastewater phosphorus loads. The sensitivity analysis compared the median RP concentration (µg/l) for the current scenario against the +/- 20% difference for diffuse arable, pasture and septic tank P sources, and wastewater P sources while holding other input values constant.*	Lines 320-236
C2.12	Additional text & Table	*The results of the parameter sensitivity analysis are presented in Table 3. Changes in point source RP loads have a greater influence on RP concentrations (µg/l) compared to diffuse sources in sub-catchment 6200 in the current scenario. A 20% increase in point source loads resulted in an 8.4% increase in RP concentrations, while a 20% reduction resulted in an 8.1% reduction in concentrations. Of the diffuse sources, arable sources had the greatest influence on RP concentration with a 20% increase yielding a 4.9% increase in concentration, while a 20% reduction resulted in a 6.5% reduction in concentrations.*	Lines 498-503 & Table 3 line 523

C2.13	Edited Figure	Figure 6 is now Figure 7 and has been updated the present median rather than mean values	Line 457
C2.13	Additional text	*Investigations of future scenarios highlighted that in the future BAU scenario (Figure 7, Pane b) median RP concentrations (µg/l) increased compared to current conditions in sub-catchments 6200, 6201 and 6205 and decreased in sub-catchments 6202 and 6206. Figure 8 for sub-catchment 6200 (and Figures S8-12) show increases in total RP loads (kg/day) in sub-catchments 6200, 6201 and 6205, while the total RP loads in sub-catchment 6202 and 6206 decreased, particularly for wastewater sources. The changes in total RP can be seen in the source apportionment between wastewater and diffuse sources, as well as the trends in climate, population and land cover change. Wastewater sources increase in sub-catchments where the population is projected to increase, while diffuse sources are expected to increase in all sub-catchments.* *In the Green Road and Fossil-Fuelled Development Extreme Precipitation scenarios, the influence of precipitation change and catchment processes are evident. Total RP loads (kg/day) are reduced in all sub-catchments in the GR ExLP scenario due to reductions in diffuse run-off. The lower likelihood of wastewater spills contributing untreated effluent to wastewater source loads are also reduced in the GR ExLP scenario. RP concentrations (µg/l) were greater in the GR ExLP scenario compared to the current scenarios in sub-catchments 6200 and 6201, despite the reductions in total RP loads in both sub-catchments (Figure 8 and Figures S8-12). We believe these concentration increases are due to the reduction in river flow volumes in the extreme low precipitation rate scenario, meaning regulating diluting functions are absent and RP concentrations increase. We are unable to investigate the influence of flows in the sub-catchments where RP concentrations decreased*	Lines 621-655

		compared to current conditions (6202, 6205 and 6206) as observed river flow volume data were not available for all sub-catchments (see SM Table 2 for more information on how surface water quality is measured absence of river flow volume data). *In the FFD ExHP scenario, increases in RP concentrations (µg/l) compared to current conditions are evident in all sub-catchment waterbodies, which is attributed to increases in total RP loads (kg/day). Increased precipitation rates increase diffuse run-off, wastewater effluent flows and the likelihood of effluent spills. For sub-catchments 6200 and 6201, despite increases in river flow volumes from increased precipitation, RP source loads into the waterbodies was greater than the dilution capacity.* *Despite 46% of the % bias observations falling within the +/- 50% acceptable model performance (Table 2), results from the goodness of fit evaluation demonstrate that the model underestimated current median RP concentrations (µg/l) at the catchment outlet in sub-catchment 6200 and the probable risk class. Simulated concentrations were more widely distributed, as compared to the observed data, as is evident in the 2% of observations within a high-risk state for simulated concentrations, compared to 0% for observed concentrations. A wider distribution in simulated RP values using a hybrid BN model was also found by Glendell et al., (2022). We concur with their considerations that both the quality and the low temporal resolutions of observed data may be responsible for this discrepancy.*	
C2.14	Edited Figures	Figures 8 and 9 have been updated to present median values and the Bayesian credible interval (Q5 and Q95) uncertainty bounds.	
C2.14	Additional supplementary Figures	We added Figures S10-13 to present Median reactive phosphorus source loads (kg/day) in waterbody sub-catchments 6201, 6202, 6205 and 6206.	Section S6 in the Supplementary Material.

C2.16	Additional text	*Despite 46% of the % bias observations falling within the +/- 50% acceptable model performance (Table 2), results from the goodness of fit evaluation demonstrate that the model underestimated current median RP concentrations (µg/l) at the catchment outlet in sub-catchment 6200 and the probable risk class. Simulated concentrations were more widely distributed, as compared to the observed data, as is evident in the 2% of observations within a high-risk state for simulated concentrations, compared to 0% for observed concentrations. A wider distribution in simulated RP values using a hybrid BN model was also found by Glendell et al., (2022). We concur with their considerations that both the quality and the low temporal resolutions of observed data may be responsible for this discrepancy.*	Lines 649-656
C2.17		*"We presented a dual representation of continuous nodes using a discretised child node to support the communication of the results using both summary statistics (median and standard deviation) available in continuous outputs and the probability of model outputs falling into agreed risk classes available in discrete variables."*	Lines 279-271
C2.19	Updated supplementary figure	We updated Table S2 by combining node name and identifier columns, removing state and discretisation columns as they are presented in Table S3 and simplifying the supporting information column to only include key information.	Table S2 of the Supplementary Material.
C2.21	Additional text	*Where supporting continuous data was available, we fitted truncated normal prior distributions by calculating the mean and standard deviation from available values. Truncated normal distributions were fitted to avoid negative values, where appropriate. Secondly, where longer data records were available, we used a built in GeNIe function to fit a custom prior distribution (histogram) to time-series data. Where available data was limited to a single deterministic value and statistical moments could not be calculated, we*	Lines 257-263

		applied scenario modelling using the diverse coupled future pathways as a best available method for representing uncertainty.	