Authors' response

Revisions (revised manuscript and list of alterations)

At the end of this document is a revised manuscript, preceeded by a list of alterations contained within it. Within this list all references made to figures, tables and page numbers relate to numbering detailed in the revised manuscript.

Response to reviewers (manuscript in discussion)

Here we provide a response to all reviewer comments. All page and line numbers, figures and tables given in this response relate to those within the manuscript "in discussion". We address all issues in the order provided by reviewer #1, followed by those provided by reviewer #2. All reviewer comments are given the prefix "R#" (in blue), and all author responses the prefix "A#" (in black).

We would like to thanks reviewer #1 for the time taken to provide such a detailed response. There are a number of points with which we agree, and have amended the manuscript in accordance with this. We feel, however, that some of the major concerns expressed by reviewer #1 (flaws in methodology, and lack of novelty in research) are unsubstantiated, perhaps stemming from some basic misinterpretations of text and tables, a misunderstanding of the study scale, and a general misconstruing of process-based modelling.

R1: "The result of the first objective, that the catchments differ in their sensitivity to climate change due to differences in soil type and nutrient transport mechanisms, is not well backed up by the results. The results section does not include any clear comparison of sub-catchment characteristics and catchment sensitivity to climate change, and neither the results nor the discussion mention other important differences between the catchments (e.g. density of tile drains). There is no acknowledgement either that with a sample of four, differences could be down to chance. Most importantly, this is also a result that has been known for decades, and this is not acknowledged. Many studies have looked at factors influencing total phosphorus (TP) export from catchments, and factors mentioned in this paper, such as soil permeability, are already taken into account in more general risk assessment tools, such as the P Index. It's then obvious that areas with higher P risk are going to be more sensitive to changes in runoff under climate change. There needs to be much more acknowledgement of this both in the intro and the discussion and conclusion, and ore/better justification for carrying out the work in the first place."
A1: We are puzzled by the first comment, whereby the reviewer begins by saying results (that sensitivity can be related back to differences in soil type) "are not well backed up", and "could be due to chance" but continues by questioning the novelty of the research, and making unsupported claims that this is information that has been "known for decades". Whilst a quick search of the literature reveals that processes involved in P export have been widely studied, little has been defined about how these processes are linked at a watershed scale (McDowell et al., 2001) and there is much still to be explored about potential catchment-wide P responses to a changing climate (Jennings et al., 2009; Kaushal et al., 2014). Multi-million pound research programmes are currently underway to develop greater understandings of these interactions (including the Macronutrient Cycles Research Programme (http://macronutrient-cycles.ouce.ox.ac.uk/), and Changing Water Cycles - specifically NUTCAT 2050 (http://nutcat2050.org.uk/)).

Where the reviewer goes on to suggest that catchment sensitivity to climate change has already been demonstrated under the P index, there seems to be some misunderstanding as to either a) the role of process-based models, or b) the appropriate scale for application of the P index. The P index was originally developed by Lemunyon and Gilbert (1993) as a tool for farmers to analyse the risk of potential P loss (due to fertiliser application) to streams from individual fields. This is not an index that is suitable for catchment scale applications. For a catchment scale application of the P index, knowledge of field-scale soil P concentrations throughout each catchment would be required. This sort of high resolution data is rarely available, and makes the P index a poor tool for large scale risk assessment applications. Furthermore, as the P index was designed specifically for easy application purposes, model inputs do not include factors that can estimate responses to specific meteorological conditions (a key feature of the manuscript); there is neither a hydrological nor a meteorological component (Reid., 2011). The "risk" output from the P index relates to that of potential P to be lost directly from soils into a watercourse; there is no consideration of in-stream processing (such as dilution, settling, desorption from the stream bed), direct inputs from sewage treatment works, or importantly the effects of aggregated risk (i.e where P contributions from a field up-stream are added to P exported from soils at downstream sites).

The P index is, of course, a very useful tool for farmers analysing the current risk of P loss from their own fields. The model cannot however be used to predict future conditions, as it relies on knowledge of current soil P conditions. To assess the risk of P loss from a catchment under a *future* climate, the P index would require an input of future soil conditions– a gap in knowledge this empirical model cannot hope to bridge.

Process based modelling, however, is designed to fill these gaps. INCA-P, which is a distributed, dynamic model, focuses on an understanding of system understandings driving phosphorus concentrations and water flow in both the terrestrial and aquatic environment. The use of this model enables us to analyse the risk to entire catchments (not just single fields) under changing conditions. By incorporating changes in climate, with those in terrestrial (soil) and aquatic (stream) conditions, the sensitivity analysis presented in this manuscript takes into account not just the current risk, but that faced by the catchment should conditions (including soil P concentrations) change. This is a key difference. The conclusion, that clay content might be used as an indicator of sensitivity is important, because geology – unlike soil P concentrations – is unlikely to change over time. Geology (or "quaternary geology" if the reviewer would prefer) is therefore a very useful risk indicator.

The shortcomings of the P index approach for projecting future catchment-scale surface water phosphorus concentrations was not originally described in the manuscript, because the field-scale P index is widely acknowledged as a field based approached, and not applicable to analyses of future change. Additional information has been included within the introduction, however, to explain the reasoning behind the use of a process based (as opposed to an empirical) model.

In response to the data being "not well backed up", tables 1 and 2 provided a detailed, quantitative cross-catchment comparison of key characteristics, with % clay composition provided on page 8073. This information is referred to in the discussion when making links between sensitivity analyses and catchment characteristics. We chose not to repeat tables of catchment characteristics in the results section, given that the information was already included here in the catchment description. For the reviewer's benefit, additional columns have been added to tables 7 and 9 (provided as tables 1 and 2 in the attached document), to directly compare sensitivity with catchment characteristics.

With regards to the possibility of results being "due to chance" – we are unsure of the reviewer's point. The comparison of hydrochemical responses over four different catchments gives a much broader understanding of catchment interactions with climate change than can be achieved from looking at a single, or pair of catchments. We are perplexed by the apparent lack of consistency from the reviewer here, given that the reviewer cites (below) three single-catchment studies and suggests these are sufficiently comprehensive as to preclude the need for any further climate change research. It is precisely because the results derived from a single catchment study can be criticized as being performed in isolation, or "due to chance", that we compared results between additional sites.

R2: "The second objective looks at climate change impacts on hydrology and water quality. The novelty here is in looking at the uncertainty within a GCM, but this is not enough in itself to justify publication, as other studies have already considered this elsewhere (e.g. Dunn et al., 2012; Fung et al., 2013). A previous paper already describes likely climate change impacts in this region wrt phosphorus (Crossman et al., 2013), and another application of INCA-P with additional climate scenarios is not, to my mind, novel enough to merit publication in itself. If the authors think this can be justified and is novel enough to be published, then the introduction needs additional detail to justify the study and put it into further context. For example, have other studies looked at the differences between uncertainty in projected future flows and TP concentrations/loads compared to the size of the projected changes?"
A2: The reviewer states concerns over the novelty of this manuscript, and writes that our aim is in "identifying uncertainty within a GCM", citing two papers that "have considered this elsewhere". This is not the focus of the manuscript. We look to identify whether in the face of an uncertain future, hydrological and chemical processes might dampen or amplify that uncertainty.

As far as we are aware, we are the first study to apply GCM uncertainty from a perturbed physics ensemble (a tool for investigating much wider ranges of plausible future climatic uncertainty which only became available within the past 6 years) to a hydrochemical impact model. Whilst it is true that the studies cited by the reviewer did look at the potential impacts of multiple GCM scenarios upon a single catchment, further novelty in our work lies in the *cross-catchment comparison* of responses to a range of climate scenarios, whereby adjacent catchments, with similar land use, demonstrate markedly different hydrochemical relationships with climate change. These important insights, both for process understanding and catchment management, could not have been determined through the assessment of uncertainty within a single catchment.

We use our results to investigate possible causes of the differences in catchment response, and whether generalised catchment characteristics might be used to predict the resilience of catchments to potential change. Having determined that %clay content might be used as an indicator, this finding could be further developed for broader scale risk applications (where data is limited), or as a tool for prioritising areas for future study (where data collection might be required). This novel research is therefore also fundamentally useful. Again, this determination would not have been made through a single catchment study. Ultimately, we must disagree with the reviewer's contention that just three single catchment studies (Dunn et al., 2012; Crossman et al., 2013; and Fung et al., 2013) are sufficient to put to rest the paradigm of climate change impacts on water quality.

R3: "There is no discussion of why process-based modelling is needed or used in the study. What is the added benefit? Would it not have been better to just compare the characteristics of the catchments and from those alone determine which was more sensitive to climate change?"
A3: For the reviewer's benefit, an additional paragraph comparing empirical and process-based models has been added to the manuscript. The reviewer's questions are confusing here and perhaps relate to a general misunderstanding of the role of process based models in the hydrological sciences. Perusal of the contents of recent issues of HESS confirms that the value of process-based modelling is widely appreciated by the hydrological community. In the specific case of this

manuscript, process-based models are required in order to calculate catchment sensitivity to climate change, and thus it is difficult to envisage how catchment characteristics alone could have been used to determine "which was more sensitive to change". The manuscript uses sensitivity derived from process-based models to demonstrate that general catchment characteristics (quaternary geology) can be used as an indicator of potential sensitivity to climate change. As this insight is a novel finding resulting from this study, the *a priori* use of geology (or other catchment characteristics) to determine sensitivity would not have been possible.

Perhaps the reviewer misunderstands the term "sensitivity" in the context of impact modelling. It refers here to *the response of water quality and hydrology to prescribed changes in future climate* (Bates et al., 2008). This was defined on page 8070. An analysis of catchment sensitivity to future change therefore requires a quantification of changes in hydrology and water quality. Differences between the hydrologic and biogeochemical response of the individual catchments are accounted for by the catchment-specific parameterization of INCA-P. Thus, variations in catchment properties (or "characteristics") were reflected in differences in INCA-P parameter values (presented in tables 1 and 2). The sensitivity was established through process-based modelling, and an assessment carried out as to dominant causative processes, finally associated back to generic catchment characteristics. The aims and objectives have been re-phrased to help clarify this.

R4: "The results and conclusions rely on using INCA-P to predict future stream flow and water quality, but no model validation was carried out, so we can have no faith in the model's predictive capacity. Model performance outside the calibration period is often significantly poorer, and the credibility of the model set-up for a given catchment must therefore be evaluated against independent data (Refsgaard and Henriksen, 2004). This test data set should test how well the model can perform the task it's intended for (different climate, in this study), problematic when looking far into the future. Refsgaard et al. (2014) provide a useful framework for this kind of modelling study."

A4: We feel that the reviewer has used the term "validation" here in a rather misleading way. The reviewer seems to suggest that by using additional independent data, the reliability of the model (in predicting future conditions) might be established. The notion of model validation in climate change impact studies was refuted over 20 years ago, in "Verification, validation and confirmation of numerical models in the Earth Sciences" (Oreskes, 1994). A match between predicted and observed data cannot prove any kind of accuracy in future predictions, because we are dealing with natural, dynamic, open systems, and no amount of independent data can account for the manner in which the system can change in unanticipated ways (Oreskes, 1994). When studying future change in natural systems we will always be left with the possibility that future observations will bring into question both our existing theory and our model calibrations. As a result of this then, as Oreskes points out, although models can be "confirmed" through a demonstration of agreement between observed and predicted data (during calibration), any sort of validation of their ability to predict future data is logically precluded, via an incomplete access to (future) natural phenomena. We thank the reviewer for drawing our attention to Refsgaard's work, who also concludes that "we cannot, in a strict sense, perform validation tests on the ability of our models to project the climate change effects since we have no data truly reflecting the future conditions" (Refsgaard et al 2014).

Thus, as we can never completely verify or validate our future projections, we are left only with the ability to calibrate our model to the best of our abilities. In summary, we disagree with the reviewer's statement that testing the model against independent data would give "more faith in the models predictive capacity"; as the model performance statistics during an observed validation period are no better an indicator of performance under a future climate than the model performance statistics during the chosen calibration period. Furthermore, recent research has shown (Larssen et al., 2007) that model performance is much greater where built on longer periods of observed data, than where shorter observed periods are used, and that it is important to use the longest time series of calibration data available (rather than to reserve additional observed data for validation). This is because where longer time periods are used, a greater range of natural phenomena are incorporated, and models can be built to predict long term trends. By calibrating the model on the largest possible range of current climate inputs, we maximise the likelihood that the model will be capable of responding to future conditions.

Ultimately, however, we stress that models are investigative tools: they can be used as a guide for management, and for prioritisation of research – but in terms of future projections, model results should be interpreted only as ranges of plausible and less plausible outcomes, given the best information currently available; results cannot be verified. Additional information on this subject has been added within the text.

R5: "Section 2.2 is lacking lots of detail on model calibration, including (i) calibration period used for each study catchment, (ii) data used for calibration, (iii) method for calibration: Graphical analysis plus manual tweaking of parameters? If so, what was the procedure followed, what performance statistics were used,. . .? (iv) Were parameters varied by sub-catchment and reach within a study catchment? (v) How many parameters therefore needed to be estimated and calibrated per study catchment? (vi) How many of these were based on some form of measured data (e.g. GIS-derived or based on literature values), how many were calibrated, but within a range derived from the literature range, and how many were purely calibrated?"
A5: We are unsure as to why the reviewer wishes to know the proportion of the total number of parameters that were calibrated using measured/observed values. This is unnecessary data, as it is most important to accurately calibrate the key parameters. Whilst obtaining measured data for 99/107 parameters within a model, a poor fit to data could still be obtained where highly sensitive parameters were not monitored. Lepisto et al (2013) and Crossman et al (2012) have determined that within INCA-P it is most important to obtain observed/accurate data for Fredulich coeffecients, soils data, and flow a parameters. This parameter information is presented in the manuscript, and INCA parameter sensitivity has now been clarified. Observed or calculated data were obtained for 30 parameters (including those identified as "key" to operation of the INCA model) and a further 6 based on literature data. The calibration procedure followed is explained in detail in section 2.2, and an overview has been added to the revised manuscript, including an outline of how the plausibility of parameter values were assessed by a) calculating P inputs [as described in the methodology and tables 1 and 2]; b) GIS and digital elevation based assessment of surface hydrological flow pathways, subcatchment size, and landuse areas; c) comparison between modelled and observed time series of flows and phosphorus concentrations [using Nash-Sutcliffe coefficients, R2 statistics, and MAE]; d) use of literature values for soil processes and e) expert judgement. The varying of all parameters by subcatchment is now included in the text for clarity. Model calibration data sources are described in detail in section 2.2, and in tables 1 and 2. Although presented in figure SI3, the period of calibration is now also repeated in the text for clarity.

R6: "To encapsulate full parametric uncertainty in the GCM, the members selected from the PPE should represent this uncertainty. However, from the description in the paper it seems that the most sensitive members were selected that still provided reasonable estimates of baseline climate. Surely the selected range should have included the least sensitive as well as the most sensitive?"
A6: We absolutely agree that the members selected from the PPE should represent the uncertainty in the GCM; this is what was done and to clarify the text has been revised accordingly. We selected 5

ensemble members from the PPE, and whilst we did indeed ensure that the "most sensitive members […] still provided reasonable estimates of baseline climate" – we also included within our selection the least sensitive ensemble members available (see table 3). Our aim was not to assess the *full* range of parametric uncertainty under a single GCM, but to assess a wider range of plausible futures than might have otherwise been considered under a multi-model ensemble (see Collins et al., 2006 for a full explanation on the benefits of PPE vs use of traditional multi-model ensembles). We include in the supplementary material (figure 1) a comparison between the range of plausible futures provided by the PPE, and those available from a more traditional multi-model approach. In short, the PPE is a more focused and deliberate approach to addressing uncertainty in climate change. Whilst the multi-model ensemble includes a range of different parameters by chance, the PPE has varied those parameters in a methodical manner. It is clear that, although some GCMs project slightly cooler futures, the selected PPE ensemble members give a much wider range of plausible future precipitation projections than would be investigated though a multi-model ensemble approach. Therefore, the members selected for use in this study enable managers to consider a broader range of conceivable scenarios. The next step would be to derive a PPE from each GCM (Collins et al., 2006); though PPEs are not yet commonly available.

R7: "For each study catchment, only one 25km2 grid cell was used to provide the climate change data (P.8077, I.13-17). It is good practice to average at least two or more RCM grid cell projections when using climate change data. In addition, later in the paper much attention is given to looking at differences between climate projections between grid cells, which is fairly nonsensical given the errors involved. I'd recommend averaging the grid cells across the whole study area and applying a single climate change scenario to the whole catchment. This would also make it easier to compare different catchment responses, as the driving climate would be the same. If you feel strongly that this is not a good way forward, then good justification needs to be given, and the authors need to show an awareness of the lack of significance of any differences in projected climate between squares when it comes to reporting results (e.g. p.8083, I17 onwards), and the discussion."

A7: It is important to consider that this is not a top-down, scenario led approach looking to formally predict the future water quality of these catchments, but a bottom-up, "stress test" of the hydrology and water quality responses of different study to different climate projections. Suggested methods for each are very different (Brown and Wilby., 2012; Nazemi and Wheat., 2014).

Given the highly localised nature of rainfall patterns in the Simcoe region (Smith., 2010), both HBV and INCA were calibrated on point data taken from catchment-specific observed weather stations (described in the text, and figure 1 of the manuscript). The reviewer suggests that we should average all 25km^2 regional climate outputs across the study area (which in total is over 2914km2) and apply a single climate scenario to all INCA models; but for the calibrated models to most accurately respond to projected climate change, they require climate data from the area closest to/covering the area under which they was calibrated.

The original GCM (at a grid resolution of 500km2) was downscaled to 25km^2 grid cells using the PRECIS regional climate model, precisely to take account of the local climate variability (due to elevation, lake effects etc.). Upscaling this detailed information back to such a large extent would introduce significantly greater errors into the modelling study. If the study were to be conducted at this scale, no regional downscaling would have been required, and would have resulted in misrepresentation of projected change within every catchment.

R8: "P.8078, I14-16: The text needs to be clearer about the very important (and quite likely invalid) assumptions involved in using delta change for bias correction, namely the assumption that the relative difference between the simulated baseline and the simulated future is realistic, despite any bias. The authors also need to make clear that this method only corrects for bias in the mean, not in the variance. Very importantly, both in the methods and in the discussion, there needs to be discussion of the fact that any potential increase in the intensity of rainfall is likely subdued using this method. This is a big source of uncertainty, particularly when looking at phosphorus, which is so affected by storm events."

A8: Delta change is the dominant method used in the US National Assessment (Hay et al., 2000) and that recommended by the IPCC (Carter., 2007). Three pages of the manuscript (8077-8079) have been devoted to the description of, and justification for use of delta change in this study. We feel that its application to the dataset has been vindicated. The claim of the reviewer that delta change is an "invalid" method is again unsubstantiated. We agree with Teutschbein and Seibert (2012) who note that this form of bias correction does not account for changes in distribution of wet and dry days, however the same can be said of linear scaling. Alternative methods such as local intensity scaling, variance scaling and power transformations each have their own drawbacks, and there is ample literature to support the use of delta change (for example Teutschbein and Seibert., 2012). We have included additional information as to the shortcomings of the delta change methods; however it should be noted that bias correction to observed station data is absolutely a necessary process within climate change applications to hydrochemcial impact models. Whilst all methods have their limitations, this does not preclude their use (Teutschbein and Seibert., 2012). On the contrary, provided the readers are made aware of the specific method used, and the corrections applied to the data (as on pages 8077 – 8079), there is no right or wrong way to bias correct the data.

R9: "Section 3.1 (Results: INCA-P model calibration): This section needs re-working, including: (i) It needs to be made clearer throughout this section what is being compared with what. Are the statistics for daily, monthly or annual means? All three are mentioned, I think, but not for every catchment. For consistency, it'd be good to give performance statistics for all time periods (daily, monthly averages and annual averages) for all sub-catchments, e.g. for the catchment outflow. It's likely that the statistics for the daily data won't be great, but if for example the performance statistics for monthly or annual TP are acceptable, then that can be used to decide over which timescale it's appropriate to discuss model output for the future period. (ii) Much of the information in the results could be put into Table 4 and the text correspondingly cut down. (iii) The results should be put into the context of 'acceptable' performance statistics from the literature (e.g. Moriasi et al., 2007), taking care to make sure that like is compared with like in terms of concentrations/loads and timescales over which the data are averaged before calculating performance statistics. (iv) This section also needs validation period statistics (v) As the point of this section is to demonstrate that the model is fit for being used to predict future conditions in the study catchments, there also needs to be some discussion of whether the right processes are operating. This is particularly important given the amount of text given over to describing catchment processes in the discussion. Some of the conclusions rely on the model having correctly simulated different flow paths, for example, so it's important to establish at this stage that the model is in fact producing realistic simulations of the different flow pathways and nutrient transport mechanisms in the different sub-catchments. (vi) Finally, as dissolved and particulate phosphorus may follow very different transport pathways to the river, it would be very interesting to consider the two separately, at least in this calibration period. This would help increase confidence that the model is performing adequately for the task in hand."

A9: All model statistics reflect monthly averages, calculated over the complete calibration time period (see response to section 5 for calibration times scales). This has been clarified in the revised manuscript. Figure 5 (of the manuscript) gives an example of the INCA model output on the daily timescale, for reference only. This is actually a rather short section, given its importance (as the reviewer also notes), and we are reluctant to relegate the information to another table. We cannot put the results in the context of "acceptable performance statistics" as there is no generically acceptable level at which a model is deemed to be "correct" (see response to point 5); every application is different, and relative only to itself. Model statistics are provided so that the reader/user can decide for themselves how much confidence they have in the results. Daily statistics are not relevant here, as only monthly averages (over 30 year time periods) are used in calculating future change. The reviewer suggestion that the calibration performance accuracy be used to determine the “timescale [over which] it’s appropriate to discuss model output for the future period” is nonsensical, as due to variance it is inappropriate to discuss future change projections in anything less than monthly averages derived over decadal periods (Carter., 2007).

We agree with the reviewer that the confirmation of the model’s ability to represent system processes is important. Equifinality is a general issue with process based modelling. Without long-term tracer studies, and other very detailed monitoring data it is difficult to ascertain whether models are correctly representing processes involved. It is however important to consider that these are multi-branched models, which have been spatially calibrated across many tributaries and a wide ranges of sites along the main stream. As a result, we can have confidence that the upstream loads travelling into each reach are accurate (based on comparison with upstream observed data), that diffuse inputs from soils are accurate (by comparing soil coefficients with other studies) and that point source inputs are accurate (from STW input reports). Whilst this in itself is not proof of the correctness of the model representation of reality, it is strong confirmative evidence that we are getting the right answers for the right reasons.

For clarification, sections 3.1 (model calibration) and 3.2 (export coefficients) have been merged in the revised manuscript, whereby the purpose of section 3.2 was to explore the modelled nutrient responses and compare those to observed data. The consistency between INCA soil export coefficients and nutrient exports derived from previous studies of these catchments (Thomas and Sevean, 1985; Winter et al., 2007; Baulch et al., 2013) gives confidence that nutrients from both diffuse and point sources are being transferred in an accurate manner.

For further clarity on model process performance, an additional analysis into model hydrological responses is presented (see Figures 2 and 3 of the attached author comments). In figure 2, model performance statistics are given separately for the rising and recession limbs of modelled verses observed hydrographs, to determine the accuracy of model responses to precipitation events in each study catchment. The same time period is analysed in each catchment during November-December 2011, during which an intense precipitation event occurred, beginning on the 27th of December. The high R2 and low model error in both rising and recession limbs of most catchments gives high confidence in the runoff simulations performed by the HBV-INCA model chain. The higher model error in the recession limb of the Pefferlaw catchment is likely due to onset of freezing in the catchment, which is not captured by the Pefferlaw at this daily timescale. The preceding event in the Pefferlaw demonstrates a more acceptable fit, with an MAE of only -1.7% on the rising limb, and 19% on the recession.

Figure 3 analyses model performance under spring melt. Analysis is performed as monthly averages during 2010, except in the case of the Whites River, where observations are rarely collected during winter, and an average of 2010 and 2011 results were required to provide sufficient data. Again high model performance is demonstrated, and gives confidence in flow pathways. Of particular note is the high seasonal accuracy of the Pefferlaw, which despite a late onset of freezing on a daily time-scale, successfully represents seasonal frozen water stores (January and February) and the transition to spring melt runoff (March). We must stress that these analyses go above and beyond those generally presented in impact modelling studies. Further verification of flow pathways would require extensive tracer studies, for which the data are simply not available. The excellent performance of

HBV in snowmelt driven catchments has been confirmed, however, through a 3-year tracer investigation in Denali National Park, Alaska (Crossman et al., 2013).

Finally, we disagree with the reviewer's suggestion that it would be "interesting to consider [dissolved and particulate phosphorus] separately" in the calibration period. Separate calibrations of dissolved and particulate P would give erroneous information. Joint calibration of total and dissolved P across all reaches increases the likelihood that we will obtain the right answer for the right reasons.

R10: "Please provide tables with all the final parameter values used, for both HBV and INCA-P, per study catchment, in the supplementary information."
A10: A table with the *key parameters* used has been provided. With over 107 parameters, in 4 different models (428 parameters) – these would be some large files. Additional information has been included here – at the reviewer's request (e.g. Plant uptake) (see table 3 of attached document).

R11: "Section 3.3 (Climate change): I recommend moving all of this to section 2.3, as this is the input to the modelling, not a result in itself. In addition, shorten the text as the key messages are somewhat lost at the moment, and rely more on data in Table 6. There also seems to be a bit of repetition between the text, table 6, Figs SI7-9. They do all show slightly different things, but probably don't merit the amount of space taken up."
A11: We disagree with the suggestion that the climate change results should be moved to the methodology. Whilst uncertainty in climate change is not the sole focus of the study, differences in projected uncertainty between catchments (under the perturbed physics ensemble) is most definitely a finding of the study; and not to be considered simply part of the methodology. Perhaps the captions were unclear on these tables and figures – these have been clarified in the revised manuscript - the information presented in tables 6 and SI7, and in figure SI8 is very different. Table 6 presents the average uncertainty across all probability levels within the CDF. Given the manner in which climate change uncertainty varies between probability levels (to different extents within different catchments) (demonstrated in SI7), table 6 gives a more complete measure of total uncertainty within a catchment (and is more appropriate for comparisons between catchments). SI8 provides information on the seasonal changes in temperature and precipitation (something that cannot be seen through the CDFs).

R12: "Throughout the paper, results are quoted too precisely(in terms of decimal places), given the errors and uncertainties. The authors also confuse significant figures and decimal places (e.g. tables 1 and 4 to 9). The number of decimal places should be reduced to 0 or 1 throughout. E.g: p.8074,l7-8: cm of snow falling given to nearest 0.1mm; reduce to nearest cm; percent changes throughout section 3.3, 3.4.1, 3.4.2 (and corresponding results tables)."
A12: Precision has been reduced accordingly

R13: "Section 3.4.2 (Water quality): This is hard to read at the moment, as too many numbers are quoted, breaking up the text. I'd suggest relying more on tables, and summarising only key results in the text. Splitting this section into sub-headings could help (e.g. total annual TP loads, monthly TP concentrations, seasonality). The 'crosscatchment range' is not very useful, it's clearer to just look at the differences between catchments. If the authors disagree, perhaps this could be pulled out of the text and summarised more."
A13: We agree that this section was rather heavy. For ease and consistency, each results section (climate change, hydrology and water quality) has been further broken up into 3 parts: i) likelihood of change, ii) seasonality of change iii) sensitivity to change. This structure has been implemented within the discussion, to help readers draw direct conclusions from the results. The numbers quoted within these sections, however, do highlight the points made. Within a PPE there is a vast quantity of

data generated, and it is a delicate balance between over-burdening the text, and providing a "predominantly qualitative description" (a criticism the reviewer raises about the site description). We are unsure what the reviewer means by " the cross-catchment range is not very useful, it's clearer to look at the differences between catchments"; given that the range is a measure of the difference. The focus of the manuscript is in comparing the range of climate change uncertainty (across the 4 catchments) with the range in hydrochemical responses. It would be difficult to justify removing these values. A clarification of the calculation used to determine the "range" has been given in the text (though this is a generally accepted mathematical term).

R14: "Section 3 (results): There is no attempt to link catchment characteristics with modelling results, despite this being one of the main objectives of the paper. A summary results table with the main differences between catchments in terms of modelling output, together with the main differences between catchments in terms of their topography, soils, etc. could be useful, plus some mention in the text."
A14: We agree that a summary results table would add clarification. Whilst tables 1 and 2 do quantitatively present the data that is referred to in the discussion, additional rows have been added to tables 7 and 9 to highlight the connection for readers (see tables 1 and 2 in attached document).

R15: "Discussion: This needs strengthening in a number of ways. It would be good if at some point it linked back to the original objectives. The first paragraph of the discussion could be deleted, as it belongs more in the introduction. Otherwise, I thought there were three main problems with the discussion: (a) There was a general lack of clarity of whether the text was referring to real observations, or simulations backed up by observations. Many of the processes mentioned in these paragraphs (e.g. loss of organic matter, macropore flow contributions and tile drainage, drainage of wetlands,. . .) aren't specifically included in INCA. These processes might indeed be important in reality, but did the modelling capture it? Need to link back to results showing it did or didn't, with a discussion of the model's limitations in relation to these key processes. Also need to discuss sooner how drainage of wetlands was taken into account in the model. (b) The discussion doesn't consider how the results fit into the wider work carried out on uncertainty in climate change, or sensitivity of different areas to P losses (even just for baseline climate). (c) The discussion doesn't consider any of the limitations or caveats of the study, of which there are many. It is crucial that these are acknowledged to not give a misleading impression of the confidence that can be placed in the results of this study."
A15: a) Additional clarity has been added to the discussion (See comment 13), though we would like to emphasize that the discussion does link directly back to the original aims. The two sub-headings, through which findings are discussed, are direct references to the study objectives stated in the manuscript introduction. Under each subheading, the implications of the results in relation to the relevant objectives are discussed – both in terms of importance to the specific study sites, and to the wider research community.

We agree that on page 8089 additional clarity was required about whether we were referring to model simulations or actual processes. This has been re-worded in the revised manuscript. We were writing of simulated outputs and simulated processes, with literature support used to demonstrate that model behaviours were plausible. All processes discussed are modelled within INCA, and changes in section 2.3 (discussion of model process clarification), along with columns in tables 7 and 9, should help with clarity here. Ultimately the aim of this section was to demonstrate that the model performance matches that suggested by observations presented in the literature, and that our modelling captured these behaviours (i.e giving further confirmation that we had obtained the right results for the right reasons).

The reviewer asks for an explanation as to how drainage of wetlands was taken into account in the model. Additional information is now provided within the text, and the additional figures (2 and 3 in the author response) should also add clarity. This was explained within section 2.3 (calibration) where it was noted that the INCA-P applications presented here used 5 landuse classes (one of which is wetlands). Calibration of landuse classes includes runoff rates, soil water storage and infiltration rates - to match observed hydrological data. As part of the calibration process therefore, the hydrological behaviour of wetlands in the Holland watershed have been adjusted to match observed hydrological outflow.

b) We disagree that the discussion did not consider how the results fit into the wider context of P sensitivity and climatic uncertainty. Page 8088 and 8089 explore the implications across all catchments where snowmelt hydrology is important. Page 8093 looks at studies of P sensitivity within catchments across Europe (e.g. the Rhine (van der Hurke, 2004)) and at a series of sites within Denmark (van Roosmalen, 2007). Finally, page 8093 discusses the wider applicability of the findings, with respect to a focus on internal process dynamics, rather than meteorological drivers.

c) We agree that it is important to clarify study limitations. Both in the presentation of methods and of results it has been made clear the accuracy of model behaviours, and that sufficient information is given to allow informed decisions by the reader as to how much confidence to place in conclusions. We have, however, drawn this information together into a summary, in a “study limitations” section, to add clarification for the reader.

R16: “Whilst the paper is reasonably well structured, the writing is not precise enough to communicate the sometimes complex concepts in a clear and transparent way (e.g. the authors confuse variance and difference, refer to model performance statistics as model coefficients, are often not clear whether they’re referring to the climate model ensemble average or members of the ensemble,. . .). I’ve highlighted quite a few examples below, in the minor comments. For methods that were used in the study, the past tense should also be used (e.g. p.8075, line 11; p.8079, line16). The present tense is confusing, sounding like a general statement of accepted science, rather than a description of methods used in this study.”
A16: We are sorry that the reviewer was confused by the description of differences between model ensemble average and individual ensemble members, and have added clarity to these. As per the reviewer’s preference, the methodology has been converted to past tense in a revised manuscript. Whilst the term “variance” was used on three occasions in place of “range” (corrected in the revised manuscript), there has also been some confusion on the part of the reviewer as the mathematical meaning of the term range (the difference between the lowest and highest value). Finally, we have clarified the manner in which we refer to “performance statistics” i.e. R^2, MAE and Nash Sutcliffe but must stress to the reviewer that it is equally common to refer to these terms as “model coefficients” (Weglarczyk, 1998; Cochrane, 1999; Saleh et al., 2000).

Minor Comments:

R1: “Introduction: Confusing absorb and adsorb several times”
A1: Thank you. The word absorb was used once in error, and has been converted to adsorb in a revised manuscript.

R2: “P.8071,l23: geological (i.e. bedrock) differences between catchments aren’t mentioned, only differences in drift and soils.”
A2: We are speaking of quaternary geology (overlies the bedrock, beneath the organic soil layer). These geological differences are described in detail. “Quaternary” has been added for clarification.

R3:”Section 2.1 (Site description): Describe available data for model calibration and testing”.
A3: Section 2.1 is a site description of the catchments. The available data used for model calibration

is described at length, both qualitatively and quantitatively under section 2.2.2 (“Model calibration”), and in tables 1 and 2. See comment 14.

R4: “P.8073, l4-28: This is an important pargraph, which currently makes for somewhat confused reading. I’d recommend summarising more, whilst keeping key information in there. Key differences between sub-catchments could be summarised, quantitatively where possible, in a table. This could then be linked to the modelling results”.
A4: It is unfortunate that the reviewer was confused here. Table 2 provides this comparison between catchments. For clarification, we have inserted a reference to this table in the suggested paragraph, and included differences in geology (or “quaternary geology”), i.e. % clay composition in tables 2, 7 and 9.

R5: “Section 2.2 (Dynamic modelling. . .): I’d suggest splitting this into (a) a description of the model; and (b) a description of the model set-up and calibration”
A5: OK

R6: “P8074 l25: the use of the word ‘parameters’ is confusing. Replace, e.g. fluxes, variables”.
A6: Agreed – changed to variables

R7. “P.8074, l25-27: confusing. I think model output timeseries are being referred to here? If so, clarify.”
A7: Clarified

R8: “P.8074, l26: soil export coefficient isn’t an output. Replace with soil erosion, this is what’s meant”.
A8: We thank the reviewer for this observation. This has been changed to “nutrient export coefficients” [from the soil]. Export coefficients represent the quantity of nutrients (or sediments) generated per unit area, per unit time (e.g. kg/ha/year). They are used for source apportionment to determine the amount of nutrients that a given landuse activity contributes to a downstream water body (McFarland and Hauck, 2001).

R9: “P.8075,l11 and l13: an individual HBV model set-up was used for each catchment, not an individual model”.
A9: Thank you. This has been altered

R10: “Model calibration: How many parameters requiring calibration does HBV have? How were these calibrated?”
A10: The method of HBV calibration has already been described (page8075, lines 6-10). A detailed model description of the parameters within HBV is provided in Crossman et al (2013) and in Saelthun (1995), and would detract here from the focus of this manuscript. Whilst it is important that the manuscript is clear about the methods used – it is important that key information is not lost by presenting details that can be obtained elsewhere.

R11: “P.8075, l20: Presumably the hydrological network was used to delineate subcatchments, rather than flow data (i.e. discharge data)? Also, how did you decide how many sub-catchments to have? On what basis?”
A11: Flow changed to “network”, and description of the use of arc-hydro to delineate catchments has been included. ArcHydro was used to develop both the flow network and to delineate subcatchments. Catchments were derived based on 1% of maximum flow accumulation (a simple rule of thumb for stream determination thresholds). As this resulted in delineation of separate catchments within a single river reach (which were not necessary for calibration), these fine scale catchments were then grouped by stream order, i.e. so that every tributary within the catchment could be individually calibrated. The decision-matrices for determining subcatchments is not relevant to the study however, and is not included in the manuscript. As INCA is an integrated model, then provided all reaches for which observed data is available are separately delineated (so as to

enable individual calibration of these areas) splitting (or merging) catchments for which data is unknown will not influence affect model accuracy.

R12: “P.8075,l20-25: refer to SI3 and tables 1 and 2.”
A12: References added.

R13: “P.8075, l.26: parameters for model calibration were ‘calculated’. This is a bit confusing, as calibration is the altering of model parameters by trial-and-error to optimise model performance.”
A13: It seems the reviewer’s understanding of the term “calibration” differs to ours. By calibration, we mean the “determination of model parameters and/or structure on basis of measurements and prior knowledge” (Janssen and Heuberger., 1995) so as to “optimise the parameter values in such a manner that the model output best fits field observation data” (Refsgaard et al. 2014). Estimation of model input parameters based on observed data is common (good) practice.

R14: “P.8076, l7-14: much of this is repeated in Table 2.”
A14: We disagree. This is a description of how the quantitative parameter values (presented in table 2) were derived (described in this section). Without the description, it would be unclear how the values were obtained.

R15: “P.8076, l14: from Table 2, I see that septic inputs were classed as inputs to nonintensive agriculture. Justify this in the text.”
A15: This has been clarified by including the detail of households with septic tanks being located in rural areas. By law, houses in urban areas must be connected to existing municipal sewage treatment systems (Ministry of Environment, 2006). Only those households situated in rural areas (and thus disconnected from the main sewage systems) may use septic tanks. As these houses are not built in the centre of corn or wheat fields (intensively farmed), on wetlands, or in dense forests - they are generally associated with “non intensive agriculture”. We did not feel that this detailed insight into the modellers’ decision making process during model set up was required.

R16: “P.8076, l.14: what about plant uptake? E.g. maximum uptake? Timing? The P budget is the key thing controlling model output, not just P inputs, so these parameters are just as important.”
A16: We agree that the P budget is important, and maximum plant uptake rates have been added to the text. The result of the difference in P budgets (export coefficients) are presented in table 5 (see response to comment 8). The reviewers opinion on the importance of P budgets (i.e inputs, outputs and processes) is a little inconsistent with their comment (no.41) – where they state that budgets are nothing more than the sum of inputs and outputs (i.e. no process interactions)? Whilst the authors agree on the importance of the P budgets, there are over 107 parameters in INCA, each of which have different degrees of influence over the model output. Common practice is to focus on those of greatest importance; previous papers have demonstrated that plant uptake is not one of these parameters (Wade et al., 2001; Lepisto et al., 2013). As noted in previous comments, details of calibration procedure for the most significant parameters have been provided in the manuscript. Determination of plant uptake values was based on previous model applications to the Simcoe catchment (Jin et al., 2013), but as direct measurements were not available, it was not originally included in table 2. We feel that a detailed discussion of all 107 model parameters would be unnecessary and unusual, to say nothing of being extremely tedious for all but the most fastidious of readers.

R17: “P.8076, l16: In Lepisto et al. (2013), the equilibrium coefficient was only mentioned in terms of a PEST-calibrated coefficient, which was then compared to lab measured values (p.56 of the report). So was PEST used for calibration? Or were their lab measured values used to decide on parameter values?”

A17: There has been a slight misunderstanding here, and the text has been clarified accordingly. The lab measured values referred to in Lepisto et al (2013) are an empirical dataset of general values for a wide range of soil and landuse types (including agriculture, forestry, water and wetlands). This dataset was used to calibrate INCA soil processes.

R18: "P.8076, l20: Were the average catchment values for EPC0 determined by area-weighting values for specific soil types, based on the area of soil in the sub-catchment?"
A18: INCA is an integrated model – so there is no "catchment averaging" done by the modellers during calibration; that is done by INCA as part of the integrated output. We entered the EPCo values for the relevant landuse type, and INCA then calculates outputs [as a function of inputs and processes] through all 5 landuses, to deliver a total P export from the respective subcatchment (into the river reach). This total export is area-weighted by landuse type, and takes into account all of the different terrestrial processes operating. What needs to be known for the input is the EPCo value for each soil (or landuse) type (see SI 1), which was provided by the laboratory data – and use of which is supported by Lepisto et al (2013), where there was correspondence between expert judgement and this data.

R19: "P.8076, l23-24: mention Fig. SI3 earlier, when model spatial set-up is described"
A19: OK

R20: "P.8076, l27: confused; re-phrase to clarify that SRES-A1B is an emission scenario; HADCM3 a GCM, and the PPE reflects parametric uncertainty in the GCM."
A20: Altered

R21: "P.8076, l28: a subset of how many members of the ensemble?"
A21: Five – this has been clarified (see comment 6)

R22: "P.8077, l1-6: this makes it sound like only two members from the ensemble were looked at (Q3 and Q10), not 5. It's then stated that Q3 and Q10 were selected because they were sensitive, then that sensitive scenarios aren't as good. This seems contradictory."
A22: Unfortunately, the reviewer misunderstood the paragraph, and it has been rephrased accordingly (see also comment 6). Five members were chosen, and these five represented as complete a range of sensitivities as was appropriate – given the need for the ensemble members to still represent baseline conditions. Members with higher sensitivities tended to have poorer representations of the regional climate – and so the highest appropriate member was Q10. Low and intermediate sensitivity members were indeed included as part of the 5.

R23: "P.8077, l25-26: bias is as important, so report that as well".
A23: This comment is inconsistent with reviewer comment 25. Delta change is, as the reviewer notes, a form of bias correction. This bias is presented in SI4 and SI5.

R24: "P.8077, l27: add 'members' after 'ensemble'"
A24: OK

R25: "P.8078, l1: delta change is a form of bias correction (as used in this paper). Therefore this needs re-phrasing, and a bit adding to clarify what bias correction method is questionable."
A25: Absolutely. This has been clarified in the revised manuscript. The section has been rephrased (see also comment 8)

R26: "P.8078, l28: little bias in simulated temperature is reported, so why was temperature than bias corrected? Bias correction introduces important errors of its own, so should only be done where the bias is more than a few degrees C."

A26: We are unsure upon what scientific basis the reviewer has determined that “a few degrees C” is the limit of necessary error for the use of a bias correction. As climate change is only expected to increase by “a few degrees C” over the next 100 years, allowing such a large bias in baseline datasets is unwise. Furthermore, if bias correction were not applied to the temperature dataset, this would result in running the INCA model using a “directly forced” temperature dataset (i.e. direct from the regional climate model output), and a “delta changed” precipitation dataset (see Lenderink et al., 2007 for the contrasts). It is good practice to be consistent with data handling.

R27:”P.8079, equations: highlight in the text that an additive change factor was used for temperature; multiplicative for precipitation.”
A27: This has been added, however it is standard practice for delta change.

R28. “P.8078, l17-18: there’s quite a lot of repetition in these two paras; merge and make more concise”
A28: OK

R29: “P.8079, l13: The text from “these time series of temperature. . .” onwards to the bottom of the section doesn’t fit in the 2.3 sub-heading; I’d recommend turning it into a new section.”
A29: OK

R30: “P.8079, line18-19: “In this way, INCA-P model deficiencies were removed”. This is incorrect: (a) model deficiencies are not removed by doing this, the model is just as deficient in the future as it is for the baseline; (b) this assumes that the deficiencies are the same for the future period as for the baseline, which is not necessarily true. For example, in the future different processes may become more or less important, which may affect model deficiencies.”

A30: Altered to “the likelihood of model deficiencies is minimised”. The reviewer’s comment is confusing, however. Parts a and b make contradictory statements (“the model is just as deficient in the future as it is for the baseline”; and “the deficiencies are [not necessarily] the same for the future period as for the baseline”). As we are using the delta change method, the model is precisely as deficient in the future as it is in the baseline. If the model over-predicts flow by 7% in the baseline, and is assumed to be just as deficient in the future, then as we are calculating percentage change in flow, this 7% over-prediction is negated, and the only changes we report are those as a response to climate alterations. To argue that model deficiencies might not be the same in the future –one could then equally say reporting of any model performance statistics during the calibration stage is pointless; as process interactions during the current period might not reflect those of the future. We must assume however, that the model performance over the current period gives some confidence in future runs (see comment 4).

R31: “P.8079, lines 19-24: sorry, I don’t quite follow here. On first reading, I understood from this that one cdf had been plotted per variable (flow, TDP, etc.), taking the variability model output using the different ensemble members to get the cdf. However, this isn’t the case as there’s one cdf plot per ensemble member. So where is the population from? Different daily values? Would be good to make a bit clearer.”
A31: Text clarified. The reviewer’s initial interpretation was correct. In the majority of figures, one CDF has been plotted per variable (flow, TDP etc) using all ensemble members to give the CDF. This is the “model ensemble CDF”. Figure 4, however, is an example of how both the ensemble average values and the uncertainty values were derived, and how the plots should be interpreted; thus it shows both the CDF ensemble (the thick black line), and the individual outputs from each ensemble (grey lines). “A” demonstrates the ensemble average projected change in temperature at the 90% probability level (thick black line). “B” demonstrates how to derive the uncertainty in temperature projections at the 90% probability level (i.e. the range between min and max ensemble members at that probability level. All values used within the CDFs represent monthly % change over the 30 year

period (daily change values are not acceptable statistics). In summary, figure 4 illustrates the PPE ensemble plot, the associated uncertainty, and how the data was derived. The methods used closely follow those of UKCP09 (http://ukclimateprojections.metoffice.gov.uk/21680).

R32: "P.8079, l26 and p.8080, l3, p.8092, l2: variance used instead of difference. Variance has a precise statistical meaning."
A32: Changed to difference

R33: "P.8080, l5-9: Delete; belongs in introduction/conclusion, but not in methods."
A33: Section moved to introduction

R34: "P.8080, l12-19: Move to methods section; not results"
A34: Moved to methods

R35: "P.8080, l19 (and throughout the text from here onwards): 'model coefficients' is confusing terminology, replace with 'model performance statistics' or similar."
A35: OK

R36: "P.8080, l21-24 and Fig. SI3: That doesn't seem justification for not including the pefferlaw to me, as there are four monitoring points in that catchment. Therefore add it to Fig. SI3 for completeness."
A36: Figure added (see figure 4 of additional material). It should be stressed that the spatial variability of accuracy within the Pefferlaw is within the range of that achieved in other catchments, and that the reason for not providing this information is simply that, with only three sites of long-term water quality monitoring for comparison (not 4 – the fourth is a gauging station where no chemical data was collected), we considered there to be no added value to presenting model accuracy for this catchment in a spatially distributed manner (as justified in the text). Unlike the Whites and Beaver, the strength of the Pefferlaw dataset lies in its temporal, rather than spatial, extent.

R37: "P.8082: this section (section 3.2) needs an introductory phrase or two to say why these are being calculated, and how this helps achieve the objectives of the study. Just by helping increase the credibility of the model? Could also be cut down."
A37: This section has now been merged with section 2.2.2 (model calibration) (see also comment 9). An introductory statement added. This section demonstrates whether model behaviours are plausible, and quantifies how much TP is being exported from each catchment (other data presented in the paper is given as % change from a baseline). This section places our study into a broader context – which is important for readers making comparisons to other studies or catchments. We also determine where the TP is coming from i.e source apportionment; which is important later in the manuscript.

R38."P.8082, l2: re-phrase as simulated average TP export coefficients for the calibration period."
A38: Adding "simulated" to the sentence is a form of reiterative redundancy - by definition all nutrient export coefficients are aerially weighted and thus are models or simulations of some kind or another.

R39: "P.8082, l6: were the previous studies of the catchments modelling or monitoring studies? Monitoring would be better."
A39: The quoted studies use monitoring data to derive their coefficients. However, see point 38: coefficients are not observed values. It is an expression of P export per unit area (an average). Whilst observed values can be used to calculate them, TP samples are rarely taken on a daily basis over a 30 year study period, and gap filling (averaging) is usually applied. A well calibrated process model, with

a daily time series output, is arguably a more accurate method than annual statistical averaging (see response to comment 40).

R40: “P.8082, l12-18: These exports from the different land uses are dependent on how the different land use classes were parameterised in INCA. To make this section relevant, it’d be good to make clear here that the point is to determine whether the simulated export fluxes are realistic, rather than presenting them as useful new results.”

A40: This comment has been taken into consideration (see comment 4) and model calibration section combined with nutrient export coefficient section. However, it is important to note that modelling of TP export coefficients are another example of how process models can be used to fill gaps in knowledge that observations cannot provide. As the reviewer rightly pointed out in comment 16, this is a function of inputs, *processes* and outputs – and the source apportionment of P exports to different landuse types could not be achieved through monitoring alone. The detailed nature of this P export data (landuse-specific) *is* new. As with any model outputs, the authors agree that model parameterisation was important. Hence the detail provided in section 2.2, and tables 1 and 2.

41. P.8082, l26-28 and p.8083, l1-3: Is this realistic? Any data? It’s just a function of the phosphorus inputs and outputs over the year (which are all very uncertain and just a function of the model parameters used), so the point of this paragraph should be to show whether the model is reasonable or not, rather than just describing something that could be unrealistic.

This comment has been addressed; with additional references used to support model behaviours. INCA demonstrates exactly what would be expected of soils responding to an input of fertiliser, and shows that the model is responding reasonably (Haygarth et al., 1998; Borling et al,. 2003).

R42: “Throughout the results sections, it would be useful if the authors, when stating results that are interesting, referred to parts of the discussion in which these interesting results were then explained and discussed in more detail (and made sure there was some discussion of them somewhere in the discussion). E.g. p.8084, l7-8. A more structured discussion with sub-headings would be needed for this to work, but I think it would make the paper tie together better.”

A42: The discussion has been restructured for clarity. We do not agree, however, with referring to sections of the discussion throughout the results. This will be confusing, and is not standard practise for this journal.

R43: “P.8084, l8: what does this mean? That 50% of the time flow increases by 23%? Is this a value from the median of the ensemble members?”

A43: No, please see table SI10. The value “23.37” is the comparison of flow projections between catchments. The paragraph summarises the data in table SI10 (which contains information from the CDF of an ensemble average of all QUMP members, and the uncertainty between members – see comment 31), showing that by 2070, there is a 50% chance that flow in the Holland and Pefferlaw will have decreased by up to 5.82 and 8.2%; and in the Whites and Beaver will have decreased by up to 29.19 and 20.35%. It comments that across the catchments, this is a large range of projected changes (23.37%) – which are significantly smaller in the Holland and Pefferlaw, and largest in the Beaver and Whites.

R44: “Section 3.4.1: re-structuring would be useful, starting with HER and SMD, and then looking at flow changes (which depend on HER and SMD). Sub-headings could help, and linking sentences describing (a) what the main change in climate change drivers is; (b) what the change in HER and SMD is, and whether this fits with the climate change drivers; (c) what the change in flow is, and whether this matches the changes in HER and SMD. It’s hard to extract this key information from the text as it is at present. Reducing reference to the cross-catchment variability would be useful (move to a table?).”

A44: This comment is confusing. Comparisons of the cross-catchment range (note: not variability) between different responses and climate drivers is the main focus of the manuscript. We have consistently followed a very clear structure for sections 3.3, 3.4.1, and 3.4.2. Each section starts with presentation of the CDF ensemble average and uncertainty for the relevant variable; continues on to discuss seasonal changes; and finishes with an analysis of sensitivity per unit change in driver. The current structure is concise, consistent, and easy to follow. We are sorry the reviewer did not see it.

R45: "P.8085, l10: I disagree that the Pefferlaw is different to the Beaver and Whites. From Table 7, the Holland is the only odd one out. Subsequent discussion needs to be altered to reflect this."
A45: Statistical significance of the difference between catchment responses cannot be analysed as there are insufficient data points. The authors are unsure, therefore, as to on what basis the reviewer states that the Holland is an "odd one out"? The original statement made is "the Holland and Pefferlaw yield the least HER in response to changes in precipitation, whilst the Beaver and Whites generate the most". The Beaver and Whites both generate a very high HER output response to changes in precipitation (almost 1:1) at 0.94 mm. The Pefferlaw generates less, at 0.87, and the Holland 0.52. Whilst the Holland certainly generates the least, the authors maintain that their original statement is correct: the Beaver and the Whites generate the most. The discussion does not need to be altered. We hope that the additional columns in tables 7 and 9 (attached) will help to clarify this.

R46: "Throughout results section: likelihood is often used, when I think the authors mean probability."
A46: This is more complicated that it at first appears. Probability is generally used by researchers in a statistical sense, involving mutually exclusive events, where all outcomes are accounted for. It has to be between 0 and 1, and all probabilities must add up to a total of 1. Statistical probability cannot be used in projecting climate change impacts, because it is impossible to account for all outcomes, due to the random and open character of natural systems. Likelihood is much weaker than probability; but essentially it gives the odds of an event given specific data. This is referred to in UKCP09 as "subjective probability", defined as "an estimate based on the available information and strength of evidence" (http://ukclimateprojections.metoffice.gov.uk/21680). The authors are happy to change all references of likelihood to "subjective probability", but would urge caution to the scientific community that this not be misinterpreted as a statistical probability, as is more commonly used.

R47: "P.8087, l17-14: It's not quite clear what's been done here. Was the daily timeseries of TP concentration, averaged over ensemble members, taken as the starting point? In Table 9 the Beaver and Whites have massive increases of 0.2 to 0.5 mg TP/l with 1mm of rainfall in one season of the year. This needs highlighting and coming back to in the discussion."
A47: Monthly averages are used, as advised by the IPCC (Carter et al., 2007). Daily changes are not appropriate in climate projections. This has been clarified in section 2.4. The method for the unit change sensitivity analysis was described in section 2.4 (originally pages 8079-8080 of the manuscript). The change in TP (mg/l) [between the baseline and future period] was divided by the change in precipitation (mm) [between the baseline and future period], giving an output of catchment response per unit change in precipitation. All values are calculated over 30 year averages, on a monthly basis. The model ensemble average was used, as this takes into account all the variations in uncertainty. Additional clarification has been provided in the methods (section 2.4).

R48: "P.8088, l21: the direction of change in projected HER and flow MUST have matched climatic drivers (precipitation, temperature), as that's what forces them. Should this just say precipitation?"
A48: We strongly disagree with this statement. The *direction of change* of HER and flow does not have to match the *direction of change* of their climatic drivers. Climatic drivers have complex interactions with a variety of processes; for example in regions where plant growth is primarily temperature limited, warmer conditions can lead to a higher growth rates, and greater primary

productivity. This can result in an increase in evapotranspiration which can offset an increase in precipitation, leading to an overall reduction in runoff. In this catchment, the interactions under discussion are during spring, and are related predominantly to ice and snow. This is a snowmelt driven catchment, and in winter and early spring, an **increase** in temperature results in a **reduction** in frozen water stores. This results in a **reduction** in spring melt. Therefore, despite an **increase** in precipitation in winter and spring, an **increased** soil moisture deficit (due to having no snow, and having removed the spring melt), results in an overall **reduction** in HER and flow. So – in spring, we have seen an increase in temperature, an increase in precipitation – but a decrease in HER and flow. The direction of change in hydrological response does not always match the direction of change in climatic drivers, when there are other significant (snowmelt) influences to consider. This is one of the reasons that detailed, process-based models are quite useful.

R49: "P.8089, l6-8: expand on this"
A49: Lines 9-24 do exactly that.

R50: "P.8089, l9: Model calibrations didn't demonstrate this, they were consistent with observations that." . .
A50: The model results do demonstrate this. As no observed time series data of soil TDP or labile P exist for these sites "consistency with observations" is not possible. It is exceptionally rare to find study areas with 30 years of observed TDP or labile P data. The literature, however, supports the model behaviours. See SI6

R51:"P.8091, l3: why?"
A51: Explained in lines 10-18.

R52: "P.8091, l8-19: again, not from my reading of the results (Pefferlaw has 0.87, which is much closer to 0.9 than it is to 0.6). All the subsequent discussion therefore needs altering."
A52: A line has been added in the discussion to clarify: "The hydrological sensitivity of the Pefferlaw is a little higher than that of the Holland, but is consistent with the difference in residence times and clay content between the two catchments (Table 2)." The hydrological sensitivity of the Pefferlaw is lower than both the Beaver and the Whites, but higher than the Holland. The difference between the Pefferlaw and the Holland is consistent with the difference in residence times and clay content (with the Pefferlaw having slightly shorter soil water residence times, and slightly higher clay content than the Holland). This is consistent with the discussion. Information has been added to tables 7 and 9 for clarity.

R53: "P.8092, l1: 0.14 mg/l, is this annual mean concentration?"
A53: It's the average coefficient of TP per unit change in precipitation over a 30 year period. See table 9.

R54: "P.8092, l3: I'm not clear what's meant by "act as a buffer to uncertainty", again line 13."
A54: To act as a "shock absorber", a "cushion", a "shield against"….

R55: "P.8092, l24-25: results not presented to back this up. Note also that soil is only a small part of geology; bedrock differences are not discussed at all. Differences in P inputs and P saturation between catchments – I can't find where that was mentioned in the results."
A55: This a puzzling comment. The manuscript goes to great length to present results on climatic inputs, and sensitivity of the catchments to the changing drivers [which the reviewer has commented upon]. In terms of differences in P inputs – these are presented in table 2. Data relating to P saturation conclusions are presented in SI6. Information derived from the climate uncertainty analysis, catchment sensitivity analysis, and data derived from the model calibration, are combined in the discussion – together with existing literature - to ascertain possible reasons for specific

hydrological and chemical responses. Additional data on clay content has been added to table 2 for clarity (and in tables 7 and 9). Differences between catchments in terms of runoff: and soil matrix flow were presented on page 8085, with significantly more HER flowing as surface runoff in the Whites and Beaver, than in the Holland and Pefferlaw.

The results in table 9 clearly indicate that timing of P export, combined with differences in runoff:matrix flow is associated with sensitivity to climate change (with the Holland and Pefferlaw being most sensitive during summer and autumn). The high P export during summer is a result of fertiliser *inputs* (shown clearly in SI6), where soil TP is highly mobile following this TP applications. There is little export in winter because the soil TP has been used up (SI6) and furthermore there is little contribution from surface runoff (page 8085). The spring/winter export in the Beaver and Whites is associated with the higher overland flow and macropore contributions, where TP is delivered directly to the streams via soils (page 8085). All of this data has now been combined and included in table 7 for clarity.

R56: “P.8094, l10: delete ‘uncertainty in’.”
A56: The sentence has been modified for clarity

R57: “P.8094, l17-18; ‘catchment sensitivity to climate uncertainty was lower. . .’; presumably should read as catchment sensitivity to climate change?”
A57: The sentence has been modified for clarity

R58: “P.8094, l10-15: not backed up by results presented here.”
A58: This again is very puzzling. These results are presented by the manuscript. The modification of table 7 should help to clarify this, and “geology” has been changed to “quaternary geology”, as requested earlier. See response to comment 55 on presentation of results

R59: “P.8095, l3-7: This doesn’t make sense; how can hydrochemical model uncertainty affect catchment sensitivity to climate change?”
A59: To address this comment, we clarify the mechanisms at work here (see also comment 3). The manuscript assesses catchment sensitivity (i.e. phosphorus and hydrological responses to climate change – see comment 3) using a process based model. The response of the INCA models to climate drivers is dependent upon their calibration accuracy, and thus uncertainty in this calibration leads inherently to uncertainty in the results. The authors went to great lengths to calibrate the models using measured data; however there are always issues such as equifinality to consider. Additional information on the caveats of the study has been added, but we would ask the reviewer to carefully consider their interpretation of the manuscript.

Comments on the Tables and figures:

R60:”Table 1: Decrease precision to just one decimal place”
A60: OK

R61: “Table 2: Add groundwater TDP concentration, parameters relating to amount and timing of plant uptake. Round catchment area to the nearest km. Re-name the first column something like ‘Parameter/Data type’, as it is not just model parameters but also input timeseries. I don’t understand the values for the first four rows of ‘hydrological characteristics’ – these are input timeseries, so what are the values? Means of some kind? For fertiliser inputs, make consistently to 1 decimal place (d.p.). Sewage inputs to 0 d.p. Define acronymms in table caption. The Beaverton is refered to as Beaver in the text.”
A61: As requested, maximum plant uptake has been included in table 2, and discussed within the manuscript. However, this is a table demonstrating sources and values for measured or calculated

input values within the calibrated INCA model, and inclusion of plant uptake values here is questionable (based on literature). Groundwater TDP is not a key model parameter (see comments 5 and 10), however it has now been added to table 2. For the reviewers' benefit, groundwater TDP data from the Provincial Groundwater Monitoring Network (http://www.ontario.ca/environment-and-energy/provincial-groundwater-monitoring-network-pgmn-data) was used to guide calibration of model parameters, and chosen inputs to the models ranged between 0.001mg/l (Beaver) and 0.007mg/l (Holland and Whites).

R62: "Table 3: Need better caption. No acronyms. Are these all the members? What are the ones in bold? What's sK? What's delta?"
A62: Here, it seems, is where the reviewer's confusion RE: comments 6 and 22 stems from. The ensemble members in bold indicate those used in the study, and clearly demonstrate a wide range of sensitivities. This has been clarified in the caption. The title has been clarified.

R63: "Table 4: Needs re-doing. Just providing locations with the best model performance statistics is not ok (cherry picking). Instead, replace with something like performance statistics for the worst and the best reaches for each study site, as well as for the catchment outflow. Please provide model performance statistics for daily data, as well as monthly and/or annual averages/loads if desired. Add in the number of observations and Nash Sutcliffe efficiency for comparability with other modelling studies. In the caption, replace 'model fit coefficients' with 'model performance statistics. Explain how the model error was calculated (difference of the means, i.e. bias or root mean squared error?)."
A63: The equation for model error has been added in the revised manuscript. Daily statistics are not appropriate for this study (see comments 9 and 31). The authors should not be accused of "cherry picking", having provided performance statistics for every reach in the catchment in Figure 6. This is a summary table of best performing reaches. At the reviewer's request, performance statistics for the whole catchment (outflow) are now also presented in this table.

R64: "Table 5: Reduce to 0 or 1 decimal places."
A64: OK

R65: "Table 6: Is the average uncertainty +/- the value given, or the width of the interval? I don't understand the units in this table (degrees C given for temperature; % for the rest). It could really help if there was a sentence in the figure caption explaining how this should be interpreted. E.g. "for the Holland sub-catchment, by 2030 precipitation simulations are +/- 19% of the ensemble average" (or whatever's correct). Decrease all to just 1 d.p."
A65: These are the units for which climate change are generally presented (^{o}C for temperature, % change for precipitation and flow). Temperature cannot be reported as a % change due to discrepancies in units (Kelvin vs Celsius). The interpretation of "average uncertainty" is explained clearly on page 8079. It is a measure of "the mean value of uncertainty across all CDF probability levels". This reason for use of this metric can be better understood by looking at table SI7. It can be seen that at different probability levels, the precipitation uncertainty in, say, the Whites, is very different (ranging from 16.4 to 24.8% in 2030). The lowest uncertainty here is at the 10% level. However, the lowest uncertainty in the Beaver projections is at the 50% level. So – it would be quite misleading to present only the ensemble average value and its corresponding uncertainty value at the 50% level. It might lead us to think that the data did not have many uncertainties within it, when in fact some catchments are very poor at conveying extreme change (10 and 90%). To better present the overall uncertainty within the different catchments, then, uncertainty was calculated at every probability level within the CDF plot, and an average taken.

The results in Table 6 don't mean that the simulations are +/- x% of the ensemble average, because this is an average of uncertainty taken across the CDF. Table 6 gives an overall uncertainty response of the PPE for each catchment and study variable, providing an effective summary for the reader. The more specific data the reviewer is looking for is provided in SI7. Table 6 is a more of a mechanism for comparison between catchment responses.

R66: "Table 8: Add dates for future periods on left hand side. Decrease all to just 1 d.p."
A66: OK

R67: "Table 9: Clarify the caption – is this averaged over one year (2030), or a 30 year period?"
A67: 2030 refers to the period 2020-2049. This was explained in the methodology. An "'s" was missing from the caption in table 9 however and has been provided.

R68: "Fig. 1: Define acronyms used in legend in the caption. Use of colour in catchment boundaries isn't good as they overlap. Annotate instead? Don't need central points of RCM squares marked. Hard to pick out sub-catchments with selected RCM squares in grey; maybe try highlighting in some other way (e.g. bold edges). Not sure what the word 'analysis' refers to in the figure caption."
A68: Figure has been adjusted for clarity

R69: "Fig. 2: Suggest deleting this figure and just giving statistics (average difference, or similar)."
A69: See comment on figure 3

R70: "Fig. 3: In caption, say what the Qs are (selected members of the PPE). The use of Q is a bit confusing, as it makes me think of quantiles, so need to be clear about this throughout."
A70: We would prefer to keep figure 2, if we are keeping figure 3. It would be inconsistent to present just the precipitation and not the temperature data. Q has been clarified in the caption

R71: "Fig. 4: Not clear what data each line is representing. Daily values? E.g. should this be interpreted as 90% of days have a temperature change less than or equal to 3.2C? An an example of how these plots should be read would be great."
A71: Figure 4 (of the manuscript) is an example of how the plots should be read. "A" demonstrates the ensemble average projected change in temperature at the 90% probability level (thick black line). "B" demonstrates how to derive the uncertainty in temperature projections at the 90% probability level i.e. the range between min and max ensemble members at that probability level. Temperature changes are monthly averages over 30 year periods.

R72: "Fig. 5: Give units."
A72: OK

R73: "Fig. 6: What is each point? Mean over whole model run? Mean of annual means?"
A73: Clarified in caption

R74: "Fig. 7: See comment on Fig. 4, and amend fig caption."
A74: OK

R75: "Figs 8 and 9: Merge into one figure. Define acronyms in figure caption."
A75: They would not be suitable as one figure – the level of detail would be too high for publication.

R76: "Fig. 10: Replace 'QUMP' with 'ensemble', or define QUMP. Are TP concentrations daily or mean monthly or seasonal? If true, say that there is one box per ensemble member."
A76: OK

R77: "All SI Figure: resolution needs increasing."
A77: OK

R78: "Fig. SI2: Define acronyms within the figure caption."
A78: OK

R79: “Fig. SI3: Why is this schematic, rather than a simple realistic map for each study catchment with the sub-catchments and reaches marked on? Why do some of the reaches appear to not connect to the main stem? Please add a scale bar for each catchment.”
A79: Scale bar added. This is a model schematic because a catchment map has already been provided. It should be noted that a) model schematics give more information as to how the model is used to represent the study catchments b) provide a more direct comparison between the availability of observed (monitoring) data, and model sub-catchment calibrations. It is more difficult to interpret this data from maps derived from satellite data. We have provided such a map for comparison (Figure 5 of the attached document) and are happy for either map to be used – but do feel the original provided greater clarification on model structure.

In relation to the “disconnection” of reaches in the schematic, this is simply a function of the INCA graphical user interface - these reaches are in fact connected to the main stem of the model.

R80: “Figs. SI4 and SI5: In the caption, need to say that Q0 to Q15 are ensemble members. Delete ‘applied to the observed data’.”
A80:OK

R81: “Fig. SI6: Which study area? Which sub-catchment? Which time period? What do the boxes represent – variability in daily labile P pools for one sub-catchment? If so, why present as boxplots rather than as a timeseries?”
A81: This is the average for intensive agricultural areas throughout all of the Holland catchment (i.e average of all subcatchments, not a single sub catchment). In both plots, monthly averages were used over the calibration period. This has been clarified in the caption.

R82: “Table SI7: Decrease to 1 or 0 decimal places.”
A82: OK

R83: “Fig. SI8 and 9, 11, 12: Define QUMP and what Q0, Q3,. . . are in the figure caption. Table SI10: This is a key table, so put in the main text, not the SI. Could be combined with Table 6. Decrease to 1 d.p. Make clear what these probability levels mean (number of days with up to this change?)”
A83: The subjective probability levels are to be interpreted as explained in section 2.3 (now section 2.4: data analysis), whereby 10%, 50% and 90% is the likelihood of change being *less than* a certain amount, having accounted for projections from all ensemble members. For example, in the Holland the temperature result at the 50% likelihood level is 2.55; this means that by 2030 there is a 50% likelihood that temperature change in any month over that 30 year average period is going to be less than 2.55^{o}C. At the 90% likelihood level the temperature value is 5.03, and means that by 2070 there is a 90% likelihood that in any month over that 30 year average period, temperature will have changed by less than 5.03^{o}C.

R84: “Table SI13: Is this the mean of the ensemble members? Is it monthly TP loads and monthly average concentrations? This is as important as table 8 in the main text; suggest moving from the SI to the main text.”
A84: Caption clarified

Reviewer 2:

We would like to thank the second reviewer for the analysis of the manuscript. The study is indeed ambitious, being the next step forward in providing practical applications of data-intensive perturbed physics ensembles for catchment management (Collins et al., 2006). We included the more intensive data in the SI, so as to enable others to replicate our methods and expand upon the work; and aiming to leave the main manuscript fully accessible to those for whom the core conclusions from the study will be generally interesting.

R1: “Is the supplementary really needed? It looks like some of the results are now presented twice, first in the main text and then in supplementary file, e.g. Figure 7 and SI7. “

A1: Information within the supplementary material is not critical to the understanding and interpretation of the manuscript. This is why it is included in the supplementary section, and not within the main text. It is true that there is some overlap between the data in figure 7, and that within SI7 – however, overall the two present quite different results. Figure 7 gives the CDF of each individual catchment, whereas the data in SI7 gives a summary of values at key percentile intervals, and the uncertainty associated with each of those percentiles. To demonstrate all of this information on one figure (different catchments and associated uncertainty) would have been too confusing – indeed, reviewer 1 was confused with just the single CDF. The table of highlighted percentiles and their uncertainties in SI7 is not critical to the paper (the numbers are quotes individually in the text), but the overall table does add context for those who wish to investigate the results more fully.

R2: “The first too pictures in supplementary are already published, and the third one is really difficult to understand. Instead would be nice to have a map which shows also location of agricultural land, wetlands and artificial areas.”

A2: The comment is understandable, but SI 1 and 2 were originally drawn by the lead author, and both have been adapted for this publication. Since their original publication, INCA-P has been modified, and an adapted schematic was required. SI3 has been re-drawn at the reviewer’s request (see figure 5 of the attached document). The absolute locations of different landuse classes are not included on the figure – the ecological land classification of Ontario is highly detailed, and does not improve clarification of the site figures. As percentage land cover has already been quantified (Table 1 of manuscript) this would not add any scientific value.

R3: “On page 8070, line 5. Eutrophication is also other harmful aspects than reduction of oxygen. In worst case it alters the whole ecosystem.”

A3: Altered to “it can result in eutrophication, where death and decay of excess algal matter leads to reduction in stream oxygen concentrations (Nicholls, 1995; Jarvie et al., 2006) which affects fish spawning and survival (Evans, 2011). In addition, the composition of algal species may be altered, and blue-green algae (e.g. cyanobacteria), can become dominant. These species may produce toxic compounds, which are harmful to terrestrial and aquatic animals (Chorus and Bartram, 1999), resulting in high water quality treatment costs (Smith, 2003).”

R4: “Only part of the PP is converted to bioavailable P, e.g. Hartikaninen et al., 2010”

A4: Absolutely. We did not intend to suggest that all PP is converted to DP. This has been rephrased for clarification.

R5: “Page 8071, line 6. Are unknown physics of climate processes really only parameter un-certainty? Not a system uncertainty etc? And parameter uncertainty comes then from model description and measurements, where measured value does not completely de-scribe what it is supposed to describe.”

A5: Absolutely, though we may be speaking at crossed-purposes. We say that parameter uncertainty stems from the unknown physics of climate systems, ie. that the unknown physics “introduces a source of uncertainty….known as parameter uncertainty”. It is not claimed that parameter uncertainty causes system uncertainty. This can be rephrased for clarification. It should be noted that these are perturbed *physics* ensembles – not just perturbed *measurement* ensembles. Here, perturbations are made not just to individual measurements of existing model functions, but to the physical structure of the model itself – so as to explore the impact of unknown physics on our certainty of projections (Collins et al., 2006). Within GCMs, the physical functions controlling the systems are generically referred to as "parameters".

R6: "In methods section would be to have description of water quality measurement, where, how often and the analysis methods."

A6: TP samples were analysed colorimetrically following digestion. These are long term records, amalgamated from a number of sources (table 2 in the manuscript) and different providers collected samples at different frequencies; this varied from twice-weekly to monthly, and included both event-based and routine sampling. These details are now included in the manuscript.

R7: "Also, some sentences of agriculture, as it covers the main land use. What is the main crop, what is the growing season. And especially, how is the climate change assumed to affect crop and growing season"

A7: Major crops in the area are alfalfa, corn (for grain), and soybeans (Statistics Canada, 2008). Livestock raised in the area are primarily poultry and cattle. The growing season for individual farms is not recorded, but are recommended by OMAFRA (as quoted in the manuscript), and it was assumed that these recommendations were followed.

INCA simulates a plant growth index, related to seasonal variation in solar radiation, which in combination with a user-specified growing season, is used to quantify nutrient uptake by plants. Should the climate become warmer, the amount of nutrients taken up could increase. Whilst a longer growing season might become possible, it is not necessarily true that it would be permitted. This would be a Ministerial decision, as Simcoe is currently an area of concern with regards to P loads, and has in place significant nutrient reduction targets.

As this is a physical assessment of the primary impacts of climate upon hydrology and water quality (and the different sensitivity of individual catchments), the secondary impacts (such as OMAFRA decisions on alterations to growing season, and subsequent choices by farmers as to changes in crops) are not incorporated; such assumptions (which would include the need for socio-economic projections of population growth) would lead to modeller bias in the interpretation of results.

R8: "Page 8074, line 5. What is the annual P? "

A8: Annual average in-stream concentrations range from 0.026mg/l in the Beaver to 0.142mg/l in the Holland. It is difficult to give comparative "true" loads, as the rivers are not monitored over the same time periods, nor at the same sampling frequency. The average annual P export (kg/km2) is given in table 5 of the manuscript.

R9: "Reference to the HBV model is missing"

A9: The version used was the Nordic HBV model (Saelthun, 1995). The reference has been added to the manuscript.

References

Borling, K. 2003. Effects of long-term inorganic fertilisation of cultivated soils. Swedish University of Agricultural Sciences, Uppsala. Doctoral Thesis.

Brown, C., and R.I. Wilby. 2012. An alternate approach to assessing climate risks. *EOS Transactions, American Geophysical Union.* 93 (41): 401-412

Carter, 2007. General guidelines on the use of scenario data for climate impact and adaptation assessment, version 2. Task Group on Data and Scenario Support for Impact and Climate Assessment (TGICA) IPCC Report

Cochrane, T.A., and Flanagan, D.C. 1999. Assessing water erosion in small watersheds using WEPP with GIS and digital elevation models. *Journal of soil and water conservation* 54 (4): 678 - 685

Collins, M., Booth, B.B.B., Harris, G.R., Murphy, J.M., Sexton, D.M.H. and Webb, M.J. 2006. Towards quantifying uncertainty in transient climate change, Clim Dynam. 27: 127 – 147

Chorus I, Bartram J, Eds (1999): Toxic cyanobacteria in water - A guide to their public health consequences. E and FN Spon, London, England

Crossman, J., Futter, M.N., and Whitehead, P.G. 2013. The significance of shifts in precipitation patterns: modelling the impacts of climate change and glacier retreat on extreme flood events in Denali National Park, Alaska. *PLOS ONE* 8 (9) e74054

Haygarth, P.M., Hepworth, L. and Jarvis, S.C. 1998. Forms of phosphorus transfer in hydrological pathways from soil under grazed grassland. Eur J Soil Sci. 49: 65 – 72

Janssen, P.H.M., and Heuberger, P.S.C. 1995. Calibration of process-oriented models. *Ecological Modelling.* 83 :55 – 66

Jarvie, H.P., Neal, C. and Withers, P.J.A. 2006. Sewage-effluent phosphorus: a greater risk to river eutrophication than agricultural phosphorus? Sci Total Environ. 360:246–53.

Jennings, E., Allott, N., Pierson, D.C., Schneiderman, E.M., Lenihan, D., Samuelsson, P., and Taylor, D. 2009. Impacts of climate change on phosphorus loading from a grassland catchment: implications for future management. *Water Reserch* 43 (17): 4316 - 4326

Jin, L., Whitehead, P.G., Baulch, H.M., Dillon, P.J., Butterfield, D., Oni, S.K., Futter, M.N., Crossman, J., and O'Connor, E.M. 2013. Modelling phosphorus in Lake Simcoe and its subcatchments: scenarios analysis to assess alternative management strategies. *Inland waters - Journal of the International Society of Limnology* 3: 207 – 220

Kaushal, S.S., Mayer, P.M., Vidon, P.G., Smith, R.M., Pennino, M.J., Newcomer, T.A., Duan, S., Welty, C., and Belt, K.T. 2014. Land use and climate variability amplify carbon, nutrient and contaminant pulses: a review with management implications. *Journal of the American Water Resources Association* 50 (3): 585-614

Larssen, T., Hogasen, T., Cosby, B.J. 2007. Impact of time series data on calibration and prediction uncertainty for a deterministic hydrogeochemical model. *Ecological Modelling.* 207: 22-33

Lemunyon, J.L, and Gilbert, R.G. 2011. The concept and need for a phosphorus assessment tool. *American Society of Agronomy* doi:10.2134/jpa1993.0483

McDowell, D., Sharpley, A., and Folmar, G. 2001. Phosphorus export from an agricultural watershed. *American Society of Agronomy* 30 (5): 1587 – 1595

McFarland, A.M.S. & Hauek, L.M. 2001. Determining nutrient export coefficients & source loading uncertainty using in-stream monitoring data. Journal of the Amer. Water Res.Assoc. 37: 223-236.

Ministry of Environment, 2006. Clean Water Act. S.O. C.22. Accessed online at www.ontario.ca/ministry-environment

Nazemi, A., and Wheater, H.S. 2014. Assessing the vulnerability of water supply to changing streamflow conditions. *EOS.* 95 (32): 288-289

Nicholls, K.H., 1995. Some Recent water quality trends in Lake Simcoe, Ontario: implications for basin planning and limnological research, Can Water Resour J. 20 (4), 213–226.

Oreskes, N., Shrader-Frechette, K., Belitz, K. 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263 (5147): 641-646

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., ... & Christensen, J. H. (2014). A framework for testing the ability of models to project climate change and its impacts. *Climatic Change*,*122*(1-2), 271-282.

Reid., K.D. 2011. A modified Ontario P index as a tool for on-farm phosphorus management. *Ontario Ministry of Agriculture, Food and Rural Affairs, Stratfor, Ontario.*

Saelthun, N. 1995. Nordic HBV Model. Norwegian Water Resources and Energy. Administration

Saleh, A., Arnold, J.G., Gassman, P.W., Hauck, L.M., Rosenthal, W.D., Williams, J.R., Mcfarland, A.M.S. 2000. Application of SWAT for the Upper North Bosque River Watershed. *Transactions of the ASAE* 43 (5), 1077-1087

Smith, V.H. 2003. Eutrophication of freshwater and coastal marine ecosystems. *Environmental Science and Pollution Research.* 10 (2): 126 – 139

Smith, G.J. 2010. Deriving spatial patterns of severe rainfall in Southern Ontario from rain gauge and radar data. *A thesis presented to the University of Waterloo.* Waterloo, Ontario.

Teutschbein, C., and Seibert, J. 2012. Bias correction of regional climate model simulations for hydrological climate-change impact studies: review and evaluation of different models. *Journal of Hydrology* 456 – 457: 12 - 29

Wade, A.J., Hornberger, G.M., Whitehead, P.G., Jarvie, H.P., and Flynn, N. 2001. On modelling the mechanisms that control in-stream phosphorus, macrophyte and epiphyte dynamics: an assessment of a new model using general sensitivity analysis. *Water Resources Research* 37 (11): 2777 – 2792

Weglarczyk, S. 1998. The interdependence and applicability of some statistical quality measures for hydrological models. *Journal of Hydrology*. 206 (1-2): 98-103

List of changes made in the manuscript (in order of responses made to comments from reviewers):

Changes made to text in response to R1 comments

- page 5 (lines 30-32) and page 6 (lines 1 – 8): justification for use of process-based models
- Addition of % clay composition component to tables 2, 7 and 9.
- All references to “geology” (as specific to this study) rephrased to “quaternary geology”
- Page 3 (lines 24-33) and page 4 (lines 1 – 3): rephrasing of study approach, aims and objectives
- page 7 (lines 1-12): explanation of calibration (and lack of validation) in climate change impact studies using process based models
- page 7 (lines 30-33) and page 8 (lines 1-5): additional information on INCA-P parametric sensitivity provided, as well as details on calibration procedure, and number of parameters calibrated
- page 7 (lines 13-15): calibration periods specified.
- Page 9 (lines 1-7): rephrasing of description of PPE member selection
- Page 7 (lines 16-17): localised nature of precipitation patterns clarified, and page 9 (line 17-18): associated justification of individual RCM grid cells given.
- Page 9 (lines 29-33), page 10 (lines 1- 3; and lines 12-18): comparison between delta change and other bias correction methods given.
- Page 12 (lines 11-13): monthly averages clarified; and equations given.
- Page 13 (lines 8 – 13) and figures 7 and 8: additional assessment of model hydrological performance during precipitation and snowmelt events provided
- Page 13 lines 14 - 33 : The originally separate section on nutrient export coefficients has been combined with model calibration results, used to demonstrate model performance in representing nutrient flow pathways
- Table 4: plant uptake and clay % added
- Precision of results altered to 1 d.p
- Results sections (3.2, 3.3 and 3.4) each further broken up into 3 parts: likelihood of change, seasonality of change, and sensitivity to change. The same structure is used within the discussion, to help readers draw direct conclusions from the results.
- Page 11 (lines 25-27): clarification of statistical meaning of the term “range”
- Additional columns added to tables 7 and 9 to highlight connection between catchment characteristics and model results
- Discussion re-structured for clarity (including sub-headings more directly relating to result and aims)
- Page 19 (lines 9 – 12) additional clarity provided as to whether referring to models or literature
- Page 7 (lines 29-31) information given on how INCA is calibrated to individual landuses (including wetlands), with figures 7 and 8 adding additional clarity
- Page 22 (lines 19-32) and page 23 (lines 1 -20): additional section on study limitations added.
- Methodology altered to past tense

- All references to “model coefficients” altered to “model performance statistics”
- All references to absorb converted to adsorb
- All references to geology altered to “quaternary geology”
- Page 8 (lines 4-5): reference added in text to table 2 and additional information on % clay composition added to figures 2, 8 and 9
- Section 2.2. presented as section 2.2.1 model setup, and 2.2.2 model calibration
- Page 6 (line 22): “parameters” changes to “variables”
- Page 6 (line 21-22): use of model output time series clarified
- All references to “export coefficient” replaces with “nutrient export coefficient”
- Page 6 (line 33): individual HBV model setup clarified
- Page 7 (line 25): flow altered to “network” and use of ArcHydro clarified
- Page 7 (lines 24-33) and page 8 (lines 1-5) references to tables 1 and 2 and to SI 3 added.
- Page 8 (lines 12 – 14): maximum plant uptake rates added
- Page 8 (lines 23-26): description of use of lab-derived EPCo and Freudlinch isotherm values clarified
- Page 7 (line 4): SI 3 references earlier
- Page 8 (lines 32) and page 9 (line 1): description of HADCM3 model rephrased
- Page 9 (line 4): number of PPE members clarified
- Page 9 (lines 1 – 7): description of PPE member selection rephrased for clarification
- Page 9 (line 27): “members” added to “ensemble”
- Page 9 (lines 29-33): use of delta change in bias correction clarified, and compared to other methods
- Page 10 (lines 27-30): delta change factors calculations for temperature and precipitation given
- Page 10 (lines 12 – 28): repetition removed and made description made more concise
- Page 11 (line 9): Text placed under “data analysis” subheading
- Page 11 (lines 11 – 13): description of model deficiencies altered
- Page 11 (lines 13 – 16): CDF interpretation clarified
- Page 11 (line 28): variance altered to difference
- Page 4 (lines 1-3): relocated from section 2.3
- Page 6 (lines 4 – 12): relocated from section 3.1
- All references to “model coefficients” rephrased to “model performance statistics”
- Pefferlaw added to figure 6
- Section on model P export coefficients now merged with section on model calibration performance.
- Page 13 (lines 14-16) use of nutrient export coefficient for increasing confidence in model behaviours explained
- Page 21 (lines 12 -14) additional information on differences between Holland and Pefferlaw
- Page 11 (lines 19-24): explanation of term “likelihood” as opposed to “probability”
- Page 11 (lines 10 - 13): use of monthly averages over 30 year time periods clarified
- Page 14 (lines 6-9): comparison of model soil TDP behaviours to literature
- Page 24 (lines 9 – 10): sentence on climate and hydrochemical future uncertainty rephrased
- Pages 22 and 23: additional information on study caveats provided

Changes made to figures and tables in response to R1 comments

- Table 1: precision reduced to 1 d.p
- Table 2: maximum plant uptake included
- Table 3: caption clarified
- Table 4: MAE equation clarified in the manuscript. Additional performance statistics for catchment outflow provided.
- Table 5: reduced to 1 d.p
- Table 8: dates added and precision reduced
- Table 9: caption clarified
- Figure 1 altered for clarification
- Figure 5: units clarified
- Figure 6 - 9: captions clarified
- Figure 19: QUMPs defined. Caption clarified
- Resolution of all SI figures increased
- SI2: acronyms defined
- SI3: altered to map rather than schematic, scale bar added
- SI7: reduced to 1 d.p
- SI8, 9, 11 and 12: QUMP defined, precision reduced. Caption clarified.
- SI 13: Caption clarified

Changes made to manuscript in response to R2 comments:

- SI 3: redrawn as map rather than schematic
- Page 2 (lines 10 – 16): additional information on alterations in algal species composition, toxicity of blooms and expenses to water quality treatments included
- Page 2 (line 20): portion of PP converted to bioavailable P clarified
- Page 3 (lines 15 – 19): clarification of system uncertainty
- Page 12 (lines 5 – 9): description of water quality measurements and analysis methods
- Page 4 (lines 18 – 19): description of major crops given (growing season previously defined)
- Page 5 (lines 26-27): annual observed P for each catchment given
- Reference to Nordic HBV model (Saelthun, 1995) provided

Flow pathways and nutrient transport mechanisms drive hydrochemical sensitivity to climate change across catchments with different geology and topography

J. Crossman*[1,2], M.N. Futter[3], P.G. Whitehead[2], E. Stainsby[4], H.M. Baulch[5], Jin. L[6], S.K. Oni[7], R.L. Wilby[8], P.J. Dillon[1]

[1] Chemical Sciences, Trent University, Peterborough, ON. Canada.
[2] Oxford University Centre for the Environment, Oxford University, Oxford. UK.
[3]Department of Aquatic Science and Assessment, Swedish University of Agricultural Science, Uppsala, Sweden
[4]Ontario Ministry of Environment, Etobicoke, ON. Canada
[5]School of Environment and Sustainability and Global Institute for Water Security, University of Saskatchewan, Saskatoon, SK. Canada
[6]Department of Geology, State University of New York College at Cortland, Cortland, NY. USA
[7]Department of Forest Ecology and Management, Swedish University of Agricultural Science, Uppsala, Sweden
[8]Department of Geography, Loughborough University, Leicestershire, UK.
*corresponding author email: jillcrossman@trentu.ca

Abstract

Hydrological processes determine the transport of nutrients and passage of diffuse pollution. Consequently, catchments are likely to exhibit individual hydrochemical responses (sensitivities) to climate change, which are expected to alter the timing and amount of runoff, and to impact in-stream water quality. In developing robust catchment management strategies and quantifying plausible future hydrochemical conditions it is therefore equally important to consider the potential for spatial variability in, and causal factors of, catchment sensitivity, as to explore future changes in climatic pressures. This study seeks to identify those factors which influence hydrochemical sensitivity to climate change. A perturbed physics ensemble (PPE), derived from a series of Global Climate Model (GCM) variants with specific climate sensitivities was used to project future climate change and uncertainty. Using the Integrated Catchment Model of Phosphorus Dynamics (INCA-P), we quantified potential hydrochemical responses in four neighbouring catchments (with similar land use but varying topographic and geological characteristics) in southern Ontario, Canada. Responses were assessed by comparing a 30 year baseline (1968-1997) to two future periods: 2020-2049 and 2060-2089. Although projected climate change and uncertainties were similar across these catchments, hydrochemical responses (sensitivity) were highly varied. Sensitivity was governed by quaternary geology (influencing flow pathways) and nutrient transport mechanisms. Clay-rich catchments were most sensitive, with total phosphorus (TP) being rapidly transported to rivers via overland flow. In these catchments large annual reductions in TP loads were projected. Sensitivity in the other two catchments, dominated by sandy-loams, was lower due to a larger proportion of soil matrix flow, longer soil water residence times and seasonal variability in soil-P saturation. Here smaller changes in TP loads, predominantly increases, were projected. These results suggest that the clay content of soils could be a good indicator of the sensitivity of catchments to climatic input, and reinforces calls for catchment-specific management plans.

Key words: Uncertainty, Phosphorus, Hydrology, Climate Change, Sensitivity

1.Introduction

Phosphorus (P) is an essential nutrient in riverine and lotic ecosystems (Jarvie et al., 2002), however when present in concentrations surplus to requirement it can result in eutrophication, where death and decay of excess aquatic plant matter leads to reduction in stream oxygen concentrations (Nicholls, 1995; Jarvie et al., 2006) which affects fish spawning and survival (Evans, 2006). High P loads can also bring about alterations in algal species composition, leading to dominance of blue-green bacteria, some of which produce toxic compounds (Chorus and Bartram, 1999), and blooms can lead to increases in water quality treatment costs (Smith, 2003). Eutrophication of many water bodies around the world is now a pressing concern making nutrient monitoring and management increasingly important.

There are numerous sources and pathways of P to rivers, which may travel either in the dissolved (DP) or particulate (PP) form. As some of the PP may be rapidly converted to bioavailable DP, the movement of both forms (total phosphorus; TP) is generally monitored. Delivery to streams may be direct through sewage effluent, or more diffuse through soils (overland flow and leaching). Leaching can be a significant P transport mechanism where the ability of soil to adsorb P is low (e.g. shortly after fertiliser applications), and where leachate rarely comes into contact with adsorption sites (e.g. through macropore flow) (Haygarth et al., 1998; Hooda et al., 1999; Borling, 2003).

Nutrient transport pathways are influenced by hydrology, which is in turn driven by topography, geology and climate. Changes in climate, therefore, have far reaching implications for the future of catchments, with increasing temperatures and altered precipitation patterns affecting the timing and magnitude of runoff and soil moisture, changing lake levels, groundwater availability and river discharge regimes (Gleik, 1989; Bates et al, 2008). It follows, therefore, that individual catchments are likely to respond to climate change in different ways (van Roosmalen *et al*., 2006). The extent to which the hydrology and nutrient concentrations of a catchment respond to alterations in climate drivers can be termed its “sensitivity” to climate change (Bates et al., 2008). In developing robust management strategies and quantifying plausible future ranges of hydrology and water quality it is important to explore possible changes in climatic and non-climatic pressures (Whitehead et al., 2009), but equally so to consider the potential for spatial variability in, and causal factors of, catchment sensitivity.

A complicating factor when determining future conditions is climatic uncertainty. This stems from a wide range of sources, including natural variability, the inability to predict future emissions of greenhouse gases, and an imprecise understanding of climate systems (modelling uncertainty) (Goddard and Baethgen, 2009). The incomplete knowledge of the physics of climate systems introduces a significant source of uncertainty into Global Climate Models, known as parameter uncertainty (Wilby and Harris, 2006; Meehl and Stocker, 2007), whereby GCMs developed by different scientists may predict various outcomes under similar forcing conditions. It is therefore important to explore a range of different plausible futures, to provide a better understanding of the vulnerability of catchments (McSweeney and Jones, 2010), including possible threshold effects. Only recently have credible techniques become available to more fully quantify the effect of physical or parametric uncertainty on future projections. These methods include multi-model- and perturbed physics ensembles (PPE). The former explores uncertainties stemming from differences in model structure by combining outputs from multiple GCMs (Collins et al., 2006). The latter alters the physical parameterisation of a single GCM, within scientifically accepted ranges, to create an array of model variants with specific climate sensitivities (UKCP, 2012). Using a PPE, alterations are made to the GCM in a systematic way, and the effects of different sensitivities on climate projections can be directly quantified (Collins et al, 2006); therefore a wider range of physically plausible future climates can be explored (McSweeney and Jones, 2010).

This study explores the underlying determinants of catchment hydrochemical sensitivity to climate change, examining if any particular catchment characteristics might be used as an indicator of that sensitivity. A comparison of hydrochemical sensitivity is made across four major sub-catchments of a large lake in Southern Ontario, Canada (Lake Simcoe), by applying five variants of the regionally downscaled Met Office Hadley Centre PPE to a range of catchments with varying characteristics, using process-based hydrochemical models. Objectives include a) comparing the hydrological and water quality sensitivity (to climate uncertainty) across catchments with different quaternary geology and topography and b) generating likelihood estimates of future water quality, having accounted for uncertainty in GCM parameters. A greater understanding of catchment sensitivity is intended to facilitate adaptive management strategies, which could look to reduce the extent of a catchment's hydrochemical response to climatic change. Such an approach would lessen reliance upon the current "project-then-act" paradigm, which relies heavily upon the accuracy of individual climate projections. The Simcoe region was chosen for study as 50 years of research has previously been undertaken into the sources and pathways of phosphorous. Agricultural practices, urbanisation, wetland drainage and sewage effluent have all been linked to increasing P levels in the lake (Evans et al., 1996; LSRCA, 2009, Jin et al., 2013), and associated with depleted oxygen concentrations and decreasing populations of trout, herring and whitefish (LSEMS, 1995).

2.Methods:

2.1 Site description

Lake Simcoe is situated in Southern Ontario, approximately 45 km north of Toronto, Canada. It drains north into the Severn River, and ultimately into Georgian Bay (Figure 1). The total catchment area is 2914 km^2 (with a lake area of 720 km^2), comprised of 21 sub-catchments. The four focus catchments (referred to in this study as Holland, Pefferlaw, Beaver and Whites) are situated in the south, and contribute some of the most significant P-inputs to the Lake (Winter et al., 2002; 2007). The combined east and west branches of the Holland form the largest catchment, with an area of 614 km^2. The Pefferlaw and Beaver are a similar size at 444 km^2 and 325 km^2; and the Whites is the smallest at 85 km^2. All catchments are characterised by a dominance of agricultural landuse (Table 1), though the Beaver and Whites have the highest coverage at over 65%. Major crops in the area are alfalfa, corn (for grain), soybeans and winter wheat (Statistics Canada, 2011). The Holland has the greatest degree of urbanisation at ~20%. There are three sewage treatment works (STW) in the Holland, two in the Beaver, and one in the Pefferlaw. Being much smaller there are no STWs within the boundaries of the Whites catchment.

A proportion of the headwaters of each catchment are located on the Oak Ridges Moraine, an important area of groundwater recharge (Johnson, 1997), characterised by steep slopes, high infiltration capacity and relatively low surface overland flow (LSRCA, 2012a). The lower reaches of each catchment flows through the Peterborough drumlin fields, Algonquin Plains and Rolling Plains (Johnson, 1997). These are characterised by low slopes, and clayey-till soils (LSRCA, 2012a). This transition in topography and quaternary geological permeability creates areas of water logging and marshlands proximal to the Lake (Miles, 2012).

Differences in catchment dominant quaternary geology leads to variations in surface overland flow and soil water residence times (table 2), and the catchments can be split into two distinct typologies; those with a high proportion of runoff, and rapid through-flow (the Beaver and the Whites) (Miles., 2012); and those transporting water more slowly through the catchment (the Holland and the Pefferlaw) (LSRCA 2012c), with less runoff, and slow rates of soil matrix flow. The Beaver and Whites have a high coverage of low-permeability clayey soils (Agriculture and Agri-Food Canada., 2010), and a high density of tile drains; characteristics which are often associated high runoff rates and macropore flow (Huang, 1995; Burt and Pinay., 2005) respectively. The saturation excess threshold of these soils is low, particularly in the northern end of the catchment, where wetlands dominate (Miles, 2012). Slower water transport in the Holland and Pefferlaw might be attributed to a higher soil storage capacity, due to the lower coverage of clayey soils, greater dominance of sandy-loams and, in the Pefferlaw, low density of tile drains and a low percentage of impervious surfaces (LSRCA, 2012b). Additionally, in the Holland, extensive seasonal drainage of wetlands is performed (LSRCA 2012c). Wetland drainage increases soil bulk density, which results in greater soil water retention capacity, and reduction of hydraulic conductivity and of overland flow (Burke, 1967; 1975; Schlotzhauer and Price, 1999). Over 50% of the Holland's former marshlands are now drained in spring for agricultural purposes, resulting in extensive modification to the catchment hydrology (Nicholls and MacCrimmon, 1975; LSRCA 2012c). Urbanisation has had little influence on surface overland flow (LSRCA 2012c), and has been attributed to township locations in the more permeable sediments of the Oak Ridges Moraine, where high infiltration capacities counteract the impact of impervious surfaces (Oni et al., 2014).

The average 30-year climate of the study regions (1968-1997) has a distinct seasonal cycle with average peak temperatures in the summer (June – August) of between 17 ^{o}C to 20 ^{o}C. Minimum temperatures occur in winter (December – February) of between -4 ^{o}C to -6 ^{o}C. Precipitation patterns are highly localised (Smith., 2010), but are generally greatest in summer (June-August) and in autumn (September-November) with between 2.9 and 3.4 mm of precipitation per day, and least in spring (March-May) with between 2.0 and 2.3 mm per day. There is a strong seasonal input of snow to the Simcoe catchment, with local meteorological stations recording an average of 92.3 cm of snow falling during winter, and 26.5 cm during spring. Annual average in-stream TP concentrations range from 0.03mg/l in the Beaver to 0.14mg/l in the Holland.

2.2. Dynamic modelling of hydrology and phosphorus using INCA-P

2.2.1 Model description

A process based, spatially distributed model was chosen for this study as it was considered more suitable for studying impacts of climate change on hydrochemistry than empirical alternatives (Adams et al., 2013). Empirical models are developed using correlative relationships, based upon a mechanistic understanding, and thus reflect current relationships between dependent and independent variables. Their applicability under future conditions is therefore questionable (Leavesley., 1994). Process-based models, however, have a greater focus on representing the underlying physical processes that describe system behaviour (Adams et al., 2013). As model parameters have a specific physical meaning, which can be defined for future altered catchment states (Bathurst and O'Connell; 1992)these models can be applied with more confidence outside the range of data under which they were developed, and hence have a greater ability to deliver credible projections under a changing climate (Leavesley., 1994).

The dynamic, process-based INCA-P model (Integrated Catchment Model of Phosphorus Dynamics) has been applied to over 40 catchments across Europe and North America (Wade et al., 2002a; Whitehead et al., 2011; Whitehead et al., 2014; Jin et al., 2013; Baulch et al., 2013 *inter alia*). It uses a semi-distributed approach to simulate the flow of water and nutrients through the terrestrial system to river reaches, differentiated by land use type. The model can simulate fully branched river networks, with unlimited numbers of tributaries and stream orders. Information flows through the model from the individual process equations, via the sub-catchment comprised of up to six land-use types, to a network of multiple reaches and tributaries. A full mass balance is imposed at each level (SI 1).

The flow of water and phosphorus through INCA is modelled through four storage zones (SI 2), using a series of detailed process equations, solved using numerical integration routines (Wade et al., 2002a,b; 2007a). By analysing phosphorus inputs from diffuse sources, sewage treatment works (STW), sediment interactions and biological processes (Wade et al., 2002a) the model estimates daily values for a range of variables both in the terrestrial and aquatic phase. Terrestrial model outputs used in this study include nutrient export coefficients, and concentrations of both soil water TDP and labile P. Aquatic outputs include flow, and concentrations of in-stream total phosphorus (TP), total dissolved phosphorus (TDP), and particulate phosphorus (PP).

INCA-P requires a daily input time series of precipitation, temperature, hydrologically effective rainfall (HER) and soil moisture deficit (SMD). Precipitation and temperature are obtained from local meteorological stations, and HER and SMD from the rainfall-runoff model Nordic HBV (Hydrologiska Byråns Vattenbalansavdelningen) (Saelthun, 1995). There are five main storage components within HBV, and the physical processes operating within and between these zones are represented through simplified mathematical expressions which can be adjusted through a series of parameters to within recommended ranges, to attain a best fit compared with observed river flow records (Oni et al., 2011). An individual HBV model set-up was used for each of the study catchments.

2.2.2 Model Calibration

Calibration periods were selected that maximise available spatial resolution and temporal longevity of observed data, and as a result, the calibration period for each catchment varied according to availability of data (SI 3). Data was not reserved for independent validation as these are open, dynamic and natural systems, and no amount of independent data can account for a system's potential to change in unanticipated ways (Oreskes, 1994); i.e. validation by independent data does not prove accuracy of future predictions any more than calibration. Again, the use of a process-based model (as opposed to empirical) gives higher confidence that models will perform adequately under future conditions (Leavesly et al., 1994; Adams et al., 2013). Lengthening the calibration period has been shown to enhance model performance and has proven the most effective method in applying water quality models to predictive studies (Larssen et al., 2007); hence the chosen method was to calibrate models using the widest possible range of conditions.

Whilst the Holland, Pefferlaw and Beaver were all calibrated for over a decade (20, 17 and 12 years respectively), the short observed record available for the Whites (2010-2012) limited calibration to three years. Intensive monitoring over 17 sites within the Whites, however, enabled a detailed calibration across the whole catchment. Given the highly localised nature of rainfall patterns within the Simcoe region (Smith, 2010), catchment-specific observed daily temperature, precipitation and flow data was used for each HBV model (PWQMN, 2009; LSEMS, 2010; LSRCA, 2010; Environment Canada, 2013; Woods, 2013 *personal communication*), and the derived HER and SMD were used to complete the daily time series required for INCA-P. For each catchment, the most proximal meteorological station data was used (Figure 1). Where station record length was insufficient, the nearest available records were used and adjusted to the local baseline conditions using bias correction techniques of Futter et al (2009).

Each of the four study catchments was subdivided into their constituent sub-catchment network using ArcHydro GIS software, and a hydrological network developed from a 2 m vertical resolution DEM (Global LandCover Facility, 2002) (SI3). Landuses were grouped into five cover classes; urban, intensive agricultural, non intensive agricultural, wetlands and forest, derived from Ecological Land Classification of Ontario data (Ontario Ministry of Natural Resources, 2007) (table 1). Parameter inputs for model calibration were then calculated. As a distributed model, INCA can be calibrated individually to each landuse, sub-catchment and river-reach within the catchment, giving extensive flexibility where conditions vary significantly within local areas. As there are over 107 parameters within the INCA model, plausibility of calibration values was assessed through a variety of methods including field measurements, GIS and digital elevation assessments, literature values, expert judgement and model performance statistics. For those model parameters which have the greatest impact on model output, it is especially important to obtain accurate calibration values. Within INCA-P these parameters include stream velocity, Freudlinch coefficients, and soil P concentrations (Crossman et al., 2013c; Lepistö et al., 2013). Details of methods used to derive values for these (and other) parameters are given below, with average values for each catchment given in Table 2.

P-inputs from fertilisers (kg P day $^{-1}$) were calculated using the methods of Wade et al (2007b) and Baulch et al (2013), from local recommended P-application rates (OMAFRA, 2009) and crop types (Statistics Canada, 2011). Inputs of P from livestock waste were calculated using current livestock assessment numbers (Statistics Canada, 2011) combined with Ontario livestock phosphorus nutrient coefficients (Bangay., 1976). Based on OMAFRA recommendations, model fertiliser inputs were extended over a 60 day period to intensively cropped areas, and a 120 day period to non-intensive agricultural lands beginning in late April (Baulch et al., 2013; Whitehead et al., 2011). Plant uptake in these agricultural areas was set to a maximum of 100 kg/ha/year, based on previous INCA-P applications to the Simcoe region (Jin et al., 2013). In addition to effluent from STWs (Table 2), a large proportion of households in rural areas contribute P loads through septic systems. Input loads were determined using methods of Scott and Winter (2006) and Paterson et al (2006), calculated as a function of annual P input per person, combined with septic system usage. Data used included LSRCA (2006) GIS information on households connected to septic systems, household population data (Statistics Canada., 2011), and the average TP load from excretion (2.6 g TP/person/day (Stephens, 2007)). As the efficiency of P removal is estimated at 57% (Whitehead et al., 2011), only 43% of the calculated septic TP load was input to the model.

Initial soil P concentrations were based on values measured by Fournier et al (1994), and soil equilibrium coefficients based on laboratory -derived equilibrium phosphorus concentration (EPC_0) and Freudlinch isotherm values for different landuse and soil types (Peltouvouri., 2006; Väänänen., 2008; Koski-Vähälä., 2001). These values were applied to the dominant soil types of each land use category within INCA (Olding et al., 1950; Agriculture and Agri-Food Canada, 2010). Average catchment values for EPC_0 ranged from 0.01 in the Beaver to 0.08 in the Holland, and are consistent with those used by Baulch et al (2013) and Whitehead et al (2011). An overview of key input variables and data sources is given in Table 2, with model structures provided in SI 3.

2.3 Climate modelling using Perturbed Physics Ensemble

The PPE used in this study is designed to quantify uncertainty in model predictions (QUMP) (McSweeney and Jones, 2010), and was developed from the HADCM3 global climate model (GCM), under the SRES-A1B scenario. Whilst ideally the widest range of ensemble member climate sensitivities would be explored, it is important that the models used continue to represent regional climatic behaviours (Collins et al., 2006). Therefore, as suggested by McSweeney and Jones (2010), a subset of five of the available 17 members were chosen, which cover the widest range of climate sensitivities, but which also accurately reproduce baseline temperature and precipitation patterns (Table 3). In the Simcoe basin, ensemble members with higher sensitivities were less representative of regional behaviours; and a maximum appropriate sensitivity of 4.0 was determined (Table 3).

The five members were dynamically downscaled by the Institute for Energy, Environment and Sustainability (IEESC,2012) to 25 km^2 grid cells using the regional climate model PRECIS (Providing Regional Climates for Impacts Studies). Complete details of the model components within PRECIS and of its application to Southern Ontario are given in Jones et al (2004), and IEESC (2012). PRECIS was run from 1968 to 2100, and daily outputs of mean temperature and total precipitation obtained for each ensemble member. Output grid cells from PRECIS were selected for analysis which covered the respective Simcoe catchments. At this resolution most catchments fall within a single grid cell; where two or more PRECIS outputs were available for a single catchment, that which was geographically closest to the catchment's observed monitoring station was selected. Individual grid cells were used for each catchment to reflect the highly localised nature of Simcoe precipitation patterns.

Whilst GCM outputs are not intended to be weather forecasts it is important that each ensemble member provides plausible climate information (Futter et al., 2009). The standard climatic baseline for Southern Ontario has been established as the 30 year period of 1968-1997 (IEESC, 2012). Therefore to establish the suitability of each scenario, monthly average comparisons for both temperature and precipitation were made between QUMP outputs and observed data over this baseline period. Although reproduction of temperature by QUMP members was highly successful (Figure 2), with an average monthly R^2 value of 0.99, precipitation was less accurate with an R^2 of only 0.23 (Figure 3). The distinctive seasonal patterns of minimum winter, and maximum summer precipitation values are well represented by ensemble members Q0, Q13 and Q15; whereas ensemble members Q3 and Q10 present inverse seasonal patterns. Whilst removing these less accurate members from the study would restrict the range of future conditions applied in the analysis, use of many of the available bias correction techniques might also be considered questionable for a PPE study. Linear scaling, local intensity scaling, variance scaling and power transformations, for example, all perturb future RCM simulations to account for biases detected during the baseline period (Lenderink et al., 2007) and in doing so would alter the future change of each ensemble member to different extents. This would render the study incapable of directly assessing impacts of climate variability (originating from internal GCM parameter uncertainty) upon catchment responses.

In contrast to the above, however, the Delta Change (Δ Change) method perturbs a time series of observed meteorological data to match projected future changes. Where there are discrepancies between baseline observed and baseline GCM data, this method assumes that the RCM is better able to simulate relative change than it is to provide absolute values (Hay et al., 2000). In the case of the PPE it also preserves the original relationship between QUMP ensemble members. Δ Change is an established technique for directly assessing impacts of climate change, and was the primary method used to generate future scenarios in the U.S National Assessment (Hay et al., 2000). It was chosen here in order to ensure plausible projections for the Simcoe region under all ensemble members, whilst incorporating the maximum range of conditions in the study. Whilst all bias correction methods have their limitations and assumptions, it should be noted that delta change specifically does not account for changes in future event frequency (Teutschbein and Seibert., 2012). The same, however, is true of the linear-scaling method. In truth, all correction algorithms assume that the same correction factor will apply under future climate conditions. Despite their various limitations, however, simulation performance is greatly improved following the correction for bias (Teutschbein and Seibert., 2012).

Δ Change was applied individually to each study catchment, using the associated gridded PRECIS outputs and observed meteorological data. Monthly Δ Change was calculated as the average monthly difference between each ensemble baseline (1968-1997) and a future 30 year period; future periods selected for comparison were 2020-2049 (the 2030s), and 2060-2089 (the 2070s). These differences were applied to the observed meteorological conditions of the baseline period, so that each month in the observed dataset underwent the same degree of change in climate as was originally demonstrated by the particular ensemble member. The standard formulae were used to determine the Δ Change for future daily temperature (Tf,daily) and future daily precipitation (Pf, daily) (Akhtar et al; 2008), where additive change factors are used for temperature, and multiplicative for precipitation:

Equation 1: $$Tf,daily = To,daily + (Tf,monthly - Tb,monthly)$$

Equation 2: $$Pt,daily = Po,daily\ x\ \frac{Pf,monthly}{Pb,monthly}$$

Where *To,daily* and *Po,daily* are the observed daily temperature and precipitation; *Tf,monthly* and *Pf,monthly* are the mean monthly QUMP simulated future temperature and precipitation;

Tb,monthly and *Pb,monthly* are the mean monthly QUMP simulated baseline temperature and precipitation. Monthly Δ change values of precipitation (%) and temperature (^{o}C) are provided in SI 4 and SI 5 for reference. The new Δ Change-adjusted temperature and precipitation time series were then passed through respective HBV catchment models to derive associated SMD and HER. Eleven time series were generated for each catchment; five ensemble members over two different future 30 year time periods, and one baseline 30 year period of observed meteorological data. Finally, these time series of temperature, precipitation, SMD and HER were used as input for the INCA-P models to generate future values of water quality (TP) and flow.

2.4 Data Analysis

Future change in variables (TP concentrations, TP loads and hydrology) were assessed as a percentage difference between 30 year baseline and 30 year future time periods. In this way, the impact of potential INCA-P model deficiencies during the calibration period were minimised, if model performance accuracy is presumed to be similar under both current and future scenarios. The ensemble cumulative distribution function (CDF) of monthly changes in each variable (temperature, precipitation, flow, and TP) was calculated to identify the likelihood of a change *being less than a certain amount*, having accounted for the projections from all ensemble members. Uncertainty was calculated as the range of values projected by ensemble members at a specific likelihood level (Figure 4), and “average uncertainty” was calculated as the mean value of uncertainty across all likelihood levels. The term “likelihood” is used in contrast to probability, where likelihood estimates give the odds of an event given specific data, and probability requires all outcomes to be accounted for. Due to the random and open character of natural systems, it is impossible to account for all possible outcomes, and therefore probability cannot be used in the statistical sense. Whilst the term “subjective probability” would also be appropriate (UKCP09, 2010), likelihood is more commonly used.

The full spectrum of catchment responses projected (both hydrological and chemical) was assessed by looking at the range of catchment results (i.e. the maximum catchment value minus the minimum catchment value). In addition, catchment sensitivity metrics were derived to assess catchment responses independent of between-catchment differences in climate drivers. The ensemble average catchment response variable (monthly change in hydrology (m^3/sec) and phosphorus concentration (mg/l)) was divided by monthly change in precipitation (mm), giving an output of “catchment response per unit change in precipitation”. A similar metric was not established for unit change in temperature due to insignificant differences in this driving variable between catchments.

Sensitivities were compared across catchments to provide an indicator as to which underlying factors might contribute to their responses.

3.Results

3.1 INCA-P model calibration

Model outputs were compared to observed flow data (PWQMN, 2009; LSEMS, 2010; LSRCA, 2010; Environment Canada, 2013; Woods, 2013 *personal communication*), and grab samples of TP. Although long term TP records were collated from a number of different sources (Table 2), all providers used similar methods of water quality analysis (colorimetric analysis following a digestion on the whole water sample). A summary of model performance statistics is provided in Table 4, an example of daily model output is given in Figure 5, and the spatial variability of model fit throughout the catchment is provided in Figure 6. All model performance statistics represent monthly averages, where model average error (MAE) is calculated as in Crossman (2013a):

Equation 3: $$\mathrm{MAE} = \left(\sum \frac{(A^t o - A^t m)}{\sum A^t o}\right) x100$$

Where $A^t o$ is the modelled value at timetep *t*, and $A^t m$ is the observed value at timestep *t*.

During calibration periods there was considerable spatial variability in TP concentration model performance statistics within each catchment (Figure 6). In general, model accuracy was greatest in catchment outflows and lower in tributaries. In the Holland the greatest flow accuracy was achieved near the outflow, with R^2 = 0.95, and MAE = -24.8%. The accuracy of TP concentrations was also highest in the main channel, with an R^2 = 0.72, and MAE = -4.4%. TP loads are well represented at the downstream extent where R^2 = 0.74. In the Pefferlaw, model performance within the main channel ranged from R^2= 0.91 to 0.65 for flow, and MAE from 14.1% to 46.3%. Near the outflow, explained variance in annual TP concentrations is fair (R^2 = 0.34) with an error of -23.4%. This need not be interpreted as a poor model fit as accuracy varied significantly with season, where MAE is as low as -13.0% in winter, but as high as -36.0% in spring and can be attributed to difficulty in representing occasional peaks in TP during March and April. Model accuracy in reproducing annual loads varies from good to excellent with an R^2 of 0.61 to 0.71.

In the Beaver, the greatest model flow accuracy was found in the upstream tributaries, where flow R^2 = 0.98, and MAE was as low as -15.3%. R^2 of TP concentrations = 0.79 and MAE was +4.4%. The explained variance for TP load was also high (up to 0.82). At the downstream extent, flow variability was simulated well, with R^2 = 0.85, though MAE was +33.5%. Although the fit to TP concentrations at the outflow was poor (R^2 = 0.05 and MAE = -58.4%) that of TP loads remained good (R^2 = 0.72).

Despite the shorter duration of observed data available for the Whites, the explained variance at the outflow for TP concentrations, flow and TP loads were 0.92, 0.29 and 0.77 respectively. Average annual flow error at this site was only 3.0%. The model does under-predict TP concentrations with an outflow MAE of -26.3%, which is lowest during autumn and spring (-6.6% and 7.6% respectively) and highest during summer (-36.2%). In headwaters and tributaries, R^2 of flow and TP loads of up to 0.90 and 0.94 were achieved. Success in modelling TP concentrations was highly spatially variable (Figure 6), with a maximum R^2 = 0.47.

In addition to the monthly model performance statistics, model responses to precipitation and seasonal snowmelt events were assessed at catchment outflows (Figures 7 and 8). High R^2 and low model errors during daily analysis of both the rising and recession limbs of hydrographs in each study catchment demonstrate the accuracy of, and give confidence in, the runoff simulations performed by the HBV-INCA model chain. Similarly, there was high seasonal accuracy of snowmelt responses (Figure 8), increasing support for the models' ability to simulate flow pathways.

TP soil export coefficients (kg TP/km^2/year) were calculated and compared with previous studies to assess the plausibility of simulated nutrient transport pathways, to determine the relative contributions of TP from each landuse type, and to identify catchments with noteworthy P export. INCA source apportionment outputs (kg TP/km^2) were divided over the years of model runs (kg TP /km^2/year). Values were highly variable between sites, with catchment averages ranging from 3.4 kg/km^2/year in the Beaver, to 13.1 kg/km^2/year in the Holland (Table 5). These findings are consistent with previous studies of these catchments (Thomas and Sevean., 1985; Winter et al., 2007; Baulch et al., 2013), although higher exports have previously been reported for the Holland (Winter et al., 2007). This could be attributed to the larger spatial domain used in this study (covering both East and West branches of the river), plus a recent decline in TP exports due to implementation of best management practices.

In each catchment, the highest exports of TP originated from areas of agricultural landuse, ranging from 3.7 kg/km^2/year (Beaver) to 16.1 kg/km^2/year (Holland). The lowest exports of TP originated from sewage treatment works, with the lowest outputs from the Whites (no output) and the greatest from the Holland (0.8 kg/km^2/year). It is notable that in the Holland exceptionally high export coefficients were found in wetlands (14.4 kg/km^2/year). This is consistent with drainage of former wetlands and use for intensive agriculture. The percentage contribution of each source to TP output from the catchment was calculated by multiplying each coefficient by landuse area, and (in the case of STWs) by calculating the load differences resulting from models run with and without STW inputs (Table 5). In the Whites, Beaver and Pefferlaw, agriculture was by far the major source of

TP (ranging from 64.3 to 69.9%). In the Holland, although agricultural sources dominate, there is also a large contribution from urban areas and STW (14.0% and 5.5% respectively).

Terrestrial model outputs indicate significant seasonal variability in soil TDP concentrations within the agricultural areas of each catchment, with large increases in summer following fertiliser additions (SI 6), and associated increases in the soil labile pool, followed by reductions in winter and spring. These simulated processes are consistent with those described by Haygarth et al (1998) and Borling (2003). The consistency of modelled nutrient and hydrological processes with existing system understanding gives confidence that catchment nutrient transport pathways are accurately represented.

3.2 Climate Change

3.2.1 Likelihood of change

Projected changes in temperature and precipitation for 10%, 50% and 90% likelihood levels indicate that changes will become progressively more extreme and more likely between the 2030s and 2070s (SI 7). Notably, there is very little difference between catchments in projected climatic drivers; by 2070 at the 50% likelihood level there is a cross catchment range of only 0.1°C (temperature) and 4.1% (precipitation). The average temperature uncertainty within all catchments was low (Table 6), at <1°C in 2030, and <2°C in 2070. Average precipitation uncertainty was higher, at <19% in 2030 and <23% in 2070. Of note is the similarity in uncertainty between catchments (Figure 9 A-D), with a cross-catchment range in average temperature uncertainty of only 3.0E-2 (2030) and 0.1 (2070) (Table 6), and cross-catchment range in average precipitation uncertainty of only 3.5% (2030) and 1.1% (2070) (Table 6).

3.2.2 Seasonality of change

Seasonal changes in climate drivers (SI 8, SI 9) demonstrate that the largest increases in temperature generally occur during winter. Average seasonal changes in temperature were similar between catchments, with a maximum cross-catchment range occurring during winter (0.2°C in 2070). The Pefferlaw was, however, consistently projected to experience the greatest winter increases (+3.0°C in 2030 and +5.9°C in 2070), and the Whites the least (+2.9°C in 2030 and 5.8°C in 2070). The largest increases in precipitation were projected to occur during winter and spring, with lower increases and modest reductions during summer and autumn. Seasonal precipitation changes varied more markedly between catchments than did temperature, with a maximum cross-catchment range during winter of 10.9% (2070). The largest changes in winter precipitation are projected to occur in the Holland (12.2%, 2030 and 26.2%, 2070), and the lowest in the Pefferlaw (5.8%, 2030 and 15.3%, 2070). The largest reductions in summer precipitation are projected for the Whites (-0.7% in 2030 and -9.2% in 2070), and the lowest for the Holland (+ 2.8% in 2030 and -2.4% in 2070).

3.3 Hydrologically effective rainfall and river flow

3.3.1 Likelihood of change

Similar to temperature and precipitation, the likelihood and extent of flow changes increased between the 2030s and 2070s (SI 10). There were, however, markedly greater differences in projected flow changes between catchments than for corresponding climatic drivers. At the 50% likelihood level there is a cross catchment range of 23.4% by 2070, where projected changes are significantly smaller in the Holland and Pefferlaw (-5.8% and -8.2%) compared to the Whites and Beaver (-29.2% and -20.4%) respectively (Figure 10). The average flow uncertainty within each catchment was also generally higher than that of temperature and precipitation, reaching up to 43.2% (2030) and 40.1% (2070) (Table 6). In contrast to climatic drivers, average uncertainty in projected flow varied considerably between catchments, with a cross-catchment range of 19.9% (2030) and 12.5% (2070), and was consistently largest in the Whites and lowest in the Pefferlaw.

3.3.2 Seasonality of change

Seasonal changes in HER and flow (SI 11, 12) demonstrate that, similar to precipitation, the largest increases occur during winter, with reductions occurring during summer and autumn. In spring, however, in contrast to the increases projected by precipitation, large reductions are projected for both HER and flow. Similar to precipitation, the maximum cross-catchment range occurred in winter, though seasonal variability in HER and flow between catchments was notably greater than that of climatic drivers, with a winter HER range of 39.0% (2030) and 68.1% (2070) and flow range of 21.7% (2030) and 46.8% (2070). Similar to precipitation, the Holland projects the largest mean seasonal increases in winter HER and flow (of up to 96.1% HER and 84.6% flow), whilst the Pefferlaw and Beaver project the smallest, with projected change as low as 28.0% (HER) and 38.0% (flow). The largest mean seasonal reductions in summer flow and HER occur in the Whites, and the smallest in the Holland and Pefferlaw.

Changes in soil moisture deficit (SMD) were also analysed. Annually, SMD increased in all catchments, though increases in the Whites and Beaver (up to 45.8%) were up to four times greater than those in the Holland and Pefferlaw (up to 11.0%). Seasonally, the largest SMD increases were projected during spring, and were consistently higher in the Beaver and Whites than those in the Holland and Pefferlaw.

3.3.3 Sensitivity to change

To determine the sensitivity of catchment HER to precipitation input, the HER generated per unit change in precipitation was calculated (Table 7). On average, the Holland and Pefferlaw yield the least HER in response to changes in precipitation, whilst the Beaver and Whites generate the most. In both the 2030s and 2070s a greater proportion of HER flowed as surface runoff in the Beaver and Whites than in the Holland and Pefferlaw (Table 7), where soil matrix flow contributions were particularly high (up to 94.3% in the Holland). The difference between catchments increased between 2030 and 2070, where proportions of surface runoff decreased in the Holland and Pefferlaw, and increased in the Beaver and Whites. Proportions of overland and subsurface flow varied by landuse type, with highest runoff proportions in saturated wetlands and urban areas, and lowest in forests and intensively farmed (and drained) agricultural lands.

3.4 Water quality

Ensemble average projections (accounting for system uncertainty) of total annual TP loads for the 2070's were projected to decrease in the Beaver and Whites by -14.7% and -13.1 % respectively, but lower reductions, and even some increases were projected for the Pefferlaw and Holland, of -1.0% and 14.7% respectively (Table 8). Projections varied considerably between catchments with a range of 14.6% (2030) to 29.4% (2070).

3.4.1 Likelihood of change

Again projected changes in monthly TP concentrations and loads at the 10%, 50% and 90% likelihood levels amplify between 2030 and 2070 (SI 13) and indicate that larger changes are more likely by 2070. The cross-catchment range of TP loads is considerable, and at 32.7% (2070 at the 50% likelihood level) is wider than that of all other variables. Similarly there is a larger difference between catchments in TP concentration changes than for corresponding climatic drivers (8.9% by 2070). Reductions in TP concentration are projected at the 50% likelihood level, and are more extreme in the Whites and Beaver than in the Holland and Pefferlaw. Similarly, at the 50% likelihood level markedly higher monthly reductions in TP loads are projected for the Whites and Beaver (-39.7% by 2070) compared to the Holland and Pefferlaw (-11.2 by 2070).

In some catchments (Holland and Pefferlaw) the average uncertainty for TP concentrations was lower than that of the corresponding climatic drivers. However, similar to flow, uncertainty varied significantly between catchments (Table 6), with a cross-catchment range of 15.0% (2030) and 16.5% (2070). In contrast, the average uncertainty for TP loads within each catchment was markedly higher than corresponding catchment drivers, flow and TP concentrations, at 30.9% in 2030, and 20.4% in

2070. Average uncertainty of TP concentrations was consistently highest in the Whites and Beaver (Table 6). Similarly, uncertainty in TP loads is highest in the Whites and lowest in the Pefferlaw (Figure 11).

3.4.2 Seasonality of change

Seasonal changes in TP concentrations demonstrate that the largest increases are projected during winter (up to +7.6%) and the largest reductions during summer and autumn (up to -26.6%) (Figure 12), and during this period are positively associated with changes in HER and flow. In spring, however, only in the Beaver and Whites did projected TP reductions (up to -9.0%) correspond with projected reductions in HER and flow. In contrast, increases in spring TP concentrations are projected for the Holland and Pefferlaw (up to 3.9%), and are negatively associated with projected hydrological changes (flow reductions). Accordingly, the spring cross-catchment range of projections was high (9.3% in 2030 and 12.9% in 2070). Ensemble average seasonal changes in TP concentration varied to a greater extent between catchments than did corresponding seasonal climatic drivers with a maximum cross-catchment range in autumn of 22.3 % (2070).

Seasonal changes in TP loads demonstrate that, similar to HER and flow, increases are projected during winter and reductions projected during spring, summer and autumn. Similar to hydrological and climatic variables, the maximum cross-catchment range occurs during winter (26.4% in 2030, and 53.2% in 2070), and minimum in spring (12.4% in 2030, and 19.1% in 2070). The degree of difference in TP loads between catchments is, however, notably higher than for all other variables, suggesting additional factors are influencing TP loads. In winter, corresponding with climatic and hydrological changes, markedly higher increases in TP load are projected for the Holland (56.2% in 2030, and 97.0% in 2070), than for the Beaver (29.8% in 2030 and 43.8% in 2070). In summer, greater reductions in TP loads are projected for the Beaver and Whites, (9.6% and -14.8% in 2030 respectively; -40.0% and -45.9% in 2070), than for the Pefferlaw and Holland (-2.2% and -6.6% in 2030 respectively; -15.1% and -20.9% in 2070). This large cross-catchment difference in seasonal balances likely accounts for a significant portion of the difference in projections of annual loads between the Pefferlaw and Holland (increases) and Beaver and Whites (reductions) (Table 8).

3.4.3 Sensitivity to change

An analysis of change in TP (mg/l) per unit change in precipitation was undertaken, to determine catchment sensitivity to climate drivers (Table 9). The Beaver and Whites had a higher sensitivity to changes in precipitation throughout the year (58.3 and 142.7 ug/l TP generated per unit of change in precipitation), compared to the Holland and Pefferlaw (22.1 and 10.6 ug/l TP per unit change in precipitation). In the Beaver and Whites the majority of changes in TP export occurred during spring and winter, in contrast to the Holland and Pefferlaw where greater changes occurred during summer and autumn.

4. Discussion

4.1 Likelihood of future changes in hydrochemistry

Climate change and variability present long term water management challenges both now, and persisting into the future. This study takes that uncertainty in our knowledge of climate systems into consideration, and results indicate that changes in climatic drivers (temperature and precipitation) become progressively more likely and extreme between the 2030s and 2070s, increasingly perturbing catchment hydrology and water quality from present conditions. Notably, although the changes in climatic drivers were similar between four neighbouring study catchments, responses in HER and flow varied considerably between sites.

With a significant seasonal snowmelt influence, catchments such as those within the Lake Simcoe basin may be very sensitive to climatic inputs (Barnett et al., 2005), and even small changes in climate could have large implications for catchment hydrology (Hamlet et al., 2005; Barnett et al., 2005), specifically with respect to the magnitude and timing of snowfall and snowmelt. The PPE projected greatest increases both in temperature and in precipitation to occur during winter, where larger quantities of precipitation previously falling as snow could in future fall as rain, likely resulting in an earlier snowmelt of reduced magnitude (Regonda et al., 2005; Crossman et al., 2013 a,b). Projected soil moisture deficits (SMD) suggest that whilst in winter these higher inputs of rain (as opposed to snow) and an earlier snowmelt will limit rises in SMD, in spring a marked SMD increase is to be expected, due in part to the earlier depletion of frozen water stores.

Seasonally, the direction of change in projected HER and flow did not always match those of climatic drivers. Although all indicated changes to be more extreme and likely between the 2030s and 2070s, spring reductions in HER and flow contrast markedly to projected precipitation increases. HER and flow are affected by a number of variables, including evapotranspiration and preceding soil moisture conditions. It is likely, therefore, that spring reductions in HER and flow are due to the increases in spring SMD, the magnitude of which was not offset by increases in spring rainfall.

The response of water quality (TP concentrations and TP loads) to climatic drivers also varied between catchments. Although, similar to climate drivers and hydrology, larger changes were more likely between 2030 and 2070, the direction of change was not consistent between catchments. Seasonally, in the Beaver and Whites the direction of change in TP concentration was consistently positively associated with that of HER and flow; whereas in the Holland and Pefferlaw a negative association was demonstrated in spring. These different hydrological and water quality responses to very similar climatic drivers are significant in terms of annual TP loads, where large annual reductions were projected for the Beaver and Whites, compared to much smaller reductions, and even some increases, projected in the Holland and Pefferlaw. Differences in hydrological flow pathways and mechanisms of nutrient delivery are a possible explanation for these different responses.

4.2 Catchment characteristics driving seasonal differences in hydrochemical responses

Model calibrations demonstrated that in every catchment TDP soil water concentrations increased in summer following fertiliser applications, which is consistent with observations of Haygarth et al (1998) where fertiliser additions and plant decomposition enriched near-surface soil stores. Research has shown that when TDP is present in excess of plant requirements, it can be transported from these surface pools to increase in-stream concentrations via overland and subsurface flow of soil water in highly fertilised organic soils, with low phosphorus sorption capacities i.e., shortly after the application of fertilisers in areas of shallow water tables e.g. wetlands (Borling., 2003). By late autumn, the surface pools of TDP may be exhausted, and lower concentrations of TDP are delivered through subsurface flows, though soils have the capacity to release some P in an attempt to re-establish an equilibrium. Generally, soil water that leaches into streams in spring and winter will be less P-enriched than it was during the summer months; the manner in which this process is represented by the models is shown in Figure SI6.

Where catchments, such as the Beaver and Whites, have significant overland and macropore flow contributions, or tile drainage (where rapid flow mobilises PP, and bypasses P-binding sites in the soils (Hooda et al., 1999)), TP can be delivered directly to the streams by erosion in spring (Heathwaite and Dils., 2000), and a positive relationship between HER and TP is maintained. Under future climates, spring flow and TP concentrations are therefore reduced simultaneously. However, where catchment pathways are dominated by soil matrix flow (Table 7), the dilute leachate enhanced by spring meltwater has a significant impact on in-stream TP concentrations (Borling., 2003), and a negative relationship between HER and TP is created. When future spring HER is reduced (and less dilute leachate delivered to streams), the in-stream TP concentrations increase. This effect may be further enhanced in the Holland and Pefferlaw catchments due to dilution of sewage treatment work contributions (point sources).

The large reductions in annual TP loads of the Whites and Beaver are projected due to large scale reductions in flow and TP concentrations in spring, summer and autumn, combined with only small scale increases in winter flow and TP concentrations. In contrast, spring flow reductions in the Holland and Pefferlaw enhance TP concentrations, which in combination with larger scale winter flow and TP concentrations (and smaller scale summer flow reductions), leads to projections of annual increases in TP loads.

The earlier thawing of frozen soils in spring under a changing climate may also influence nutrient flow pathways. In the Holland and Pefferlaw, earlier thawing was projected to lead to an increased proportion of subsurface flow, enhancing interactions with the soil matrix, and may be an additional causal factor behind the dilution of the spring and winter leachate. In the Beaver and Whites, however, with large macropores and tile drainage, soil freeze-thaw is projected to have less impact on infiltration (as observed by Harris., 1972). In addition, as water movement through the soil profile in the Beaver and Whites is rapid, any changes in subsurface flow generated during an earlier spring melt would have a limited effect on soil water concentrations reaching the stream. A final consideration is the drainage of wetlands (polders) in the Holland River during spring (Burke., 1967, 1975), where artificial drainage of soils can significantly increase levels of P lost through subsurface leaching (McDowell et al., 2001). This practice might therefore contribute to the negative HER associated with TP during spring, where the addition of artificial drainage waters with high TP concentrations will in future be less diluted by flows with a lower spring melt contribution.

4.3 Impact of catchment sensitivity compared to future climate uncertainty

In all catchments, uncertainty in temperature is low (across the range of QUMP scenarios generated by the PPE), and similar between catchments. As is common amongst climate change studies, uncertainty in precipitation is higher (Hawkins and Sutton., 2011; Maskey et al., 2004), though remains relatively consistent between sites. Importantly, hydrological (HER and flow) uncertainty is considerably higher than corresponding climatic drivers, and varies more significantly between catchments. The catchment-specific response can be defined by the quantity of HER generated per unit of precipitation input (Table 7). This relative index of HER sensitivity to climate change normalises the catchment response from differences in original driving forces (precipitation). On average, the Beaver and Whites catchments demonstrated a greater HER sensitivity to precipitation change (>0.9mm) than the Holland and Pefferlaw (<0.6mm). This sensitivity might be attributed to shorter soil water residence times in the Beaver and Whites, whereby more rapid soil drainage through macropores and tile drains results in prompt responses of soil moisture content to changes in precipitation; and the dominance of overland flow results in greater exposure to evaporation and rising temperatures. As a result, reductions in summer and autumn precipitation have a greater impact upon SMD in the Beaver and Whites (as indicated by model results) and thus upon HER and flow. Similarly, reductions in spring-melt waters within these catchments have a greater impact on SMD, as soils with a dominant overland flow component would previously have responded most rapidly and significantly to snowmelt input (Dunne and Black., 1971).

In contrast, in the Holland and Pefferlaw projected changes in SMD are lower, and the impact of a changing climate upon hydrology less extreme, due to a higher soil storage capacity, longer soil water residence times, and a greater proportion of water directed through the soil matrix. Slow passage of waters through soils in the Holland and Pefferlaw reduces its response rate to precipitation changes and meltwater reductions, and its exposure to evaporation (Dunne and Black., 1971; van den Hurk et al., 2004), meaning soil moisture deficits can be recharged with a lesser affect upon HER and stream flow. The hydrological sensitivity of the Pefferlaw is higher than that of the Holland, but is consistent with the difference in soil water residence times and clay content between the two catchments (Tables 2 and 7).

The sensitivity of TP concentrations to changing climate also varies considerably between catchments, with a greater response of TP per unit change in precipitation in the Whites and Beaver (up to 142.7 ug/l) than in the Holland and Pefferlaw (up to 22.1 ug/l). In Simcoe this variance between catchments appears to be associated with flow pathways and timings of nutrient export which may act as a buffer to uncertainty within some catchments. For instance, in the Holland and Pefferlaw, the majority of changes in TP export are associated with changes in future summer and autumn precipitation (Table 9), despite there being smaller absolute changes in water volumes during these periods. This is likely due to the high mobility of soil TP in the Holland and Pefferlaw at this time of fertiliser applications, with high initial P soil concentrations, and limited P sorption capacities. As the uncertainty of future climate change is lowest during this period, these are the catchments for which the lowest TP uncertainty is calculated. During the periods of high climate uncertainty (winter and spring), TP mobility is low, and the effects of the climate change uncertainty upon water quality are buffered. In the Beaver and Whites, a much larger proportion of the changes in TP export occurs during winter and spring. This is likely due to the high runoff rates and macropore flow, which facilitates movement of TP directly to the stream during all seasons. Climate uncertainty is highest during the large projected spring and winter changes, therefore uncertainty in future TP concentrations are higher in these catchments. Given that nutrient loads are directly derived from a combination of flow and TP concentrations, it follows that the uncertainty from each propagates up into the load estimations. As such, uncertainty of future TP loads is also greatest in the Whites.

It is evident then that the future certainty in both hydrology and water quality depends not only upon climatic inputs (between which there is little difference in these catchments), but also largely upon the relative sensitivity of a catchment to changing drivers. Results from this study suggest that catchment sensitivity is strongly influenced by quaternary geology, seasonal P-inputs and P-saturation thresholds. Sensitivity to climatic drivers was highest in clayey catchments, dominated by overland flow, with little influence of P-saturation (Beaver and Whites); it was lowest in sites with greater contributions from soil matrix flow, and where seasonal variability in soil P-saturation had a marked influence on water quality (Holland and Pefferlaw). It is likely that these characteristic drivers of sensitivity will be applicable to other catchments. For instance, van den Hurke (2004) determined that future inter-annual variability in wintertime runoff in the Rhine was smallest in catchments with high storage capacities, whilst van Roosmalen et al (2007) determined from a range of sites in Denmark with differing quaternary geology, that sites with clayey soils responded to precipitation change with the highest increases in stream discharge and overland flow proportions. The consistency of driving forces behind catchment sensitivity to climate change requires further investigation, but indicates that the clay content of soils might be used as an indicator of catchment hydrochemical sensitivity to climate change.

4.4 Study limitations and further research:

By analysing outputs from a range of physical structures of a single global climate model (a perturbed physics ensemble), this study explores the sensitivity of catchments to uncertainty in climate system behaviour and parameter values of global climate models. Whilst this enables an examination of a wider range of plausible future projections than could be investigated through a multi-model ensemble approach, the PPE used is derived only from the HADCM3 model. A wider range of plausible futures could be generated by perturbing the physics of additional GCMs (Collins et al., 2006). Currently, however, PPEs are available for only a limited number of GCMs.

As discussed previously, the choice of GCM bias correction method impacts both climatic and hydrochemical outputs. The Δ Change method was selected as it preserves the impact of perturbing the physical structure of the climate model (i.e. it maintains the difference between PPE ensemble members). This method, however, does not account for potential changes in distribution of wet and dry days, and outputs from this study should not be used in assessments of future flood or drought frequency.

Finally, parametric uncertainty within both INCA-P and the rainfall-runoff model HBV might also influence model outputs. In this study, care was taken to accurately calibrate the most sensitive parameters. Within HBV these have previously been identified as parameters associated with partitioning precipitation into runoff and soil moisture (Seibert, 1997; Uhlenbrook et al., 1999). The assessment of hydrological responses to individual storm events and seasonal snowmelt events (Figures 7 and 8) enhance confidence in runoff partitioning within these model applications. Within INCA-P, sensitive parameters have been identified as Freudlich soil coefficients and stream flow velocities (Crossman et al., 2012; Lepistö et al., 2013). Laboratory and field values specific to each catchment were used for these parameters. The similarity of model process responses (including soil water TDP responses to fertiliser inputs) to those in the literature, and of model soil TP export coefficients to those calculated by Winter et al (2007), gives confidence in model representation of reality. While models were calibrated using these measured and calculated values, there are over 107 parameters within INCA, and parametric uncertainty is an issue that should be considered (Wade et al., 2001; Starrfelt and Kaste., 2014). Although beyond the scope of this study, future work will explore the extent to which parametric uncertainty might influence assessments of catchment sensitivity.

Whilst there are a range of uncertainties associated with hydrochemical modelling, process based modelling does provide the best method of understanding catchment responses to possible future conditions (Jin et al., 2013). These results provide a useful basis for the development of catchment-specific management approaches. The overall outcome of this study indicates that management strategies might focus on internal processes dynamics that define the extent of hydrochemical responses to changes in climatic input, and as such reinforces the call for catchment-specific management plans. In the Holland and Pefferlaw, where the majority of P export appears to be derived from leachate during summer and autumn, a reduction in labile P concentrations and an increase in soil sorption capacity would be beneficial. It has been shown that soil P levels can take over 20 years to fully recover from extensive periods of excess fertiliser applications (McCollum., 1991; McLauchland., 2006), but may gradually be achieved through methods such as avoidance of nutrient applications in excess of plant requirements (Whitehead et al., 2011), and establishing nutrient vulnerable zones where use is no longer required. In the Beaver and Whites, where the majority of P export appears to be derived from surface runoff and soil erosion, a more appropriate strategy might include efforts to reduce rates of surface runoff from fertilised fields, through injecting fertiliser directly to the root zone, constructing vegetated riparian areas (Sharpley et al., 2001), restricting fertiliser applications to periods of higher soil moisture deficit (Westerman and Overcash., 1980; Sharpley et al., 2001), and limiting tile drainage.

5. Conclusion

Accounting for climate modelling uncertainty, this study suggests that in the Beaver and Whites, large scale reductions in annual TP loads could result from reductions in flow and TP concentrations in spring, summer and autumn, combined with minor increases in winter. In contrast, much smaller reductions and even some increases in annual TP loads are expected in the Holland and Pefferlaw, due to much smaller scale reductions in summer and autumn flow and TP concentrations, combined with larger scale increases in winter, and enhanced spring TP concentrations. These findings indicate that all changes will become more extreme and more likely between 2030 and 2070.

Uncertainty in future projections of climate change had different impacts on uncertainty of future hydrological and water quality projections within each catchment. This impact depended upon a) catchment characteristics which determine the sensitivity of a catchment to climate drivers and b) the degree of projected climate change within each catchment. Influential catchment characteristics include quaternary geology, seasonal P-inputs and P-saturation thresholds. Catchment sensitivity to climatic drivers was relatively high where catchments had large proportions of overland flow and a direct association between TP concentrations, loads and HER was established. Catchment sensitivity was lower in catchments with larger proportions of soil matrix flow, and where seasonal variability in soil P-saturation influenced water quality; here internal catchment processes had a greater impact on hydrology and water quality.

The different ranges of uncertainty demonstrated by the catchments indicate that in water quality management it is important to consider every catchment as a distinct hydrological unit. Effectiveness of management strategies will vary depending on the dominant nutrient sources and P transport mechanisms in each area. Significantly, although the hydrology and water quality of the Whites and Beaver Catchments are more sensitive to changes in climate, the majority of projections under the PPE indicate reductions in TP concentrations and loads. Although the Holland and Pefferlaw are less sensitive to change, the dominant response is an increase in annual TP loads; as such the latter catchments might be seen as priority target areas for nutrient management.

6. Acknowledgements

The authors would like to thank Jennifer Winter at the Ontario Ministry of Environment, for her invaluable scientific inputs and support throughout the project. Thanks also to the Hadley Centre, for generating the original PPE datasets, and to Professor Huang and associates at the University of Regina, for generating the regionally downscaled PRECIS datasets of Southern Ontario (Huang et al., 2012). Finally, thanks to Dave Woods and Eavan O’Connor, for collection of water quality and flow monitoring data in the Lake Simcoe catchments. Funding for this study was provided by the Ontario Ministry of Environment (grant number 55-58313).

7. References

Adams, H.D., Williams, A.P., Chonggang, X., Rauscher, S.A., Jiang, X., and McDowell., N.G. 2013. Empirical and process-based approaches to climate-induced forest mortality loads. Front.Plant.Sci. doi: 10.3389/fpls.2013.00438

Agriculture and Agri-Food Canada. 2010. Soil Landscapes of Canada version 3.2. *Soil Landscapes of Canada Working Group*

Akhtar, M., Ahmad, N., and Booij, M.J. 2008. The impact of climate change on the water resources of Hindukush-Karakorum-Himalaya region under different glacier coverage scenarios, J Hydrol 355: 148 - 163

Bangay, G.E. 1976. Livestock and poultry wastes in the Great Lakes Basin. Environmental Concerns and Management Issues, Environment Canada Social Science Series No. 15.

Barnett, T.P. and Adam, .J.C., Lettenmaier, D.P. 2005. Potential impacts of a warming climate on water availability in snow-dominated regions, Nature. 438(7066): 303-309. doi:10.1038/nature04141. PubMed: 16292301

Bates, B.C., Kundzewicz, Z.W., Wu, S. and Palutikof, J.P. (Eds.). 2008. Climate Change and Water. Technical Paper VI of the Intergovernmental Panel on Climate Change. IPCC Secretariat, Geneva, Switzerland.

Bathurst., J.C., and O'Connell., P.E. 1992. Future of distributed modelling: the Systeme Hydrologique Europeen. *Hydrological Processes.* 6: 265 - 277

Baulch, H.M., Futter, M.N., Whitehead, P.G., Woods, D.T., Dillon, P.J., Butterfield, D.A., Oni, S.K., Aspden, L.P., O'Conner, E.M. and Crossman, J. 2013. Phosphorus dynamics across intensively monitored subcatchments in the Beaver River, J Inland Wat. 1 (1) 187 - 206

Borling, K. 2003. Effects of long-term inorganic fertilisation of cultivated soils. Swedish University of Agricultural Sciences, Uppsala. Doctoral Thesis.

Burke, W. 1967. Principles of drainage with special reference to peat. Irish Forestry 24, 1–7.

Burke, W. 1975. Aspects of the hydrology of blanket peat in Ireland. Hydrology of marsh-ridden areas. Proceedings of the Minsk symposium, June 1972. IAHS Studies and Reports in Hydrology, 19. Paris: Unesco Press, 171–82

Burt, T.P. and Pinay, G. 2005. Linking hydrology and biogeochemistry in complex landscapes, Prog Phys Geog. 29 (3): 297 – 316

Chorus I, Bartram J, Eds (1999): Toxic cyanobacteria in water - A guide to their public health consequences. E and FN Spon, London, England

Collins, M., Booth, B.B.B., Harris, G.R., Murphy, J.M., Sexton, D.M.H. and Webb, M.J. 2006. Towards quantifying uncertainty in transient climate change, Clim Dynam. 27: 127 – 147

Crossman, J., Futter, M.N., Oni, S.K., Whitehead, P.G., Jin, L., Butterfield, D., Baulch, H.M. and Dillon, P.J. 2013a. Impacts of climate change on hydrology and water quality: future proofing management strategies in the Lake Simcoe watershed, Canada, J Great Lakes Res. 39 (1): 19-32

Crossman, J., Futter, M.N. and Whitehead, P.G. 2013b. The significance of shifts in precipitation patterns: modelling the impacts of climate change and glacier retreat on extreme flood events in Denali National Park, Alaska, PLOS ONE. 8 (9) e74054

Crossman, J., Whitehead., P.G., Futter, M.N., Jin, L., Shahgedanova, M., Castellazzi, M., Wade, A.J. 2013c. The interactive responses of water quality and hydrology to changes in multiple stressors, and implications for the long-term effective management of phosphorus. *Science of the Total Environment* 454-455: 230-244

Dunne, T. and Black, R.D. 1971. Runoff processes during snowmelt, Water Resour Res. 7 (5): 1160 – 1172.

Environment Canada. 2013. Water survey of Canada. www.wsc.ec.gc.ca . Accessed January 2014

Evans, D.O., Nicholls, K.H., Allen, Y.C. and McMurtry, M.J., 1996. Historical land use, phosphorus loading, and loss of fish habitat in Lake Simcoe, Canada, Can J Fish Aquat Sci. 53, 194–218.

Evans, D.O. 2006. Effects if hypoxia on scope-for-activity of lake trout: defining a new dissolved oxygen criterion for protection of lake trout habitat. *Aquatic Research and Development Section, Applied Research and Development Branch, Ministry of Natural Resources*

Fournier, R.E., Morrison, I.K. and Hopkin, A.A. 1994. Short range variability of soil chemistry in three acidic soils in Ontario, Canada, Commun Soil Sci Plan. 25 (17-18): 3069 - 3082

Futter, M.N., Forsius, M., Homberg, M. and Starr, M. 2009. A long-term simulation of the effects of acidic deposition and climate change on surface water dissolved organic carbon concentrations in a boreal catchment. Hydrol Res 402(3):291–305.

Gleik, P.H. 1989. Climate change, hydrology, and water resources, Rev Geophys. 27 (3), 329–344

Global LandCover Facility. 2002. Digital Elevation Model. Earth Science Data Interface. Available: http://www.landcover.org/ Accessed 2013 March 1

Goddard, L. and Baethgen, W. 2009. Better estimating uncertainty with realistic models, EOS. 90 (39): 343-344

Hamlet, A.F., Mote, P.W., Clark, M.P. and Lettenmaier, D.P. 2005. Effects of temperature and precipitation variability on snowpack trends in the Western United States, Am Met Soc. 18: 4545-4561

Harris, A.R. 1972. Infiltration rate as affected by soil freezing under three cover types. *American Society of Agronomy*. 36 (3): 489 - 492

Hawkins, E. and Sutton, R. 2011. The potential to narrow uncertainty in projections of regional precipitation change, Clim Dynam. 37: 407 - 418

Hay, L.E., Wilby, R.L. and Leavesley, G.H. 2000. A comparison of delta change and downscaled GCM scenarios for the three mountainous basins in the United States, J Am Water Resour As. 36 (2): 387-397

Haygarth, P.M., Hepworth, L. and Jarvis, S.C. 1998. Forms of phosphorus transfer in hydrological pathways from soil under grazed grassland. Eur J Soil Sci. 49: 65 – 72

Heathwaite, A.L. and Dils, R.M. 2000. Characterising phosphorus loss in surface and subsurface hydrological pathways. Sci Total Environ. 25: 523 - 538

Hooda, P,S., Moynaghm M., Svoboda, F.I., Edwards, A.C., Anderson, H.A. and Sym, G. 1999. Phosphorus loss in drainflow from intensively managed grassland soils, Am Soc Agron. 28 (4): 1235 - 1242

Huang, M. 1995. Old and subsurface water contributions to storm runoff generation in flat, fractured, clayey terrain. Department of Civil and Environmental Engineering, University of Windsor. *PhD Thesis*

[IEESC] Institute for Energy, Environment and Sustainable Communities. 2012. Producing High-Resolution (25km by 25km) Probabilistic Climate Change Projections over Ontario Using UK PRECIS. Report submitted to the Ontario Ministry of Environment (Eds Huang, G.H., Wang, X.Q., Lin, Q.G., Yao, Y., Cheng, G.H., Fan, Y.R., Li, Z., Lv, Y., Han, J.C., Wang, S., Suo, M.Q., Dong, C., Chen, J.P., Chen, X.J., Zhou, X). Data available online at: http://env.uregina.ca/moe/

Jarvie, H.P., Neal, C., Williams, R.J., Neal, M., Wickham, H., Hill, L.K., Wade, A.J., Warwick, A. and White, J. 2002. Phosphorus sources, speciation and dynamics in a lowland eutrophic chalk river, Sci Total Environ. 282 – 283(175):203.

Jarvie, H.P., Neal, C. and Withers, P.J.A. 2006. Sewage-effluent phosphorus: a greater risk to river eutrophication than agricultural phosphorus? Sci Total Environ. 360:246–53.

Jin, L., Whitehead, P.G., Baulch, H.M., Dillon, P.J., Butterfield, D., Oni, S.K., Futter, M.N., Crossman, J. and O’Connor, E.M. 2013. Modelling phosphorus in Lake Simcoe and its subcatchments: scenario analysis to assess alternative management strategies, J Inland Wat. 3: 207 - 220

Johnson, F.M. 1997. The Landscape Ecology of the Lake Simcoe Basin, Lake Reserv Manage. 13 (3): 226 – 239. DOI: 10.1080/07438149709354313

Jones, R.G., Noguer, M., Hassel, D.C., Hudson, D., Wilson, S.S., Jenkins, G.J. and Mitcehll, J.F.B. 2004. Generating high resolution climate change scenarios using PRECIS, Met Office Hadley Centre. Exeter, UK. 1-40

KMK Consultants (2004) Schomberg Water Pollution Control Plant Class Environmental Assessment. Environmental Study Report. Prepared for The Regional Municipality of York, Ontario

Koski-Vähälä, J. 2001. Role of resuspension and silicate in internal phosphorus loading. Dissertation in Limnology. Department of Limnology and Environmental Protection; Department of Applied Chemistry and Microbiology. University of Helskini. ISBN: 952-10-0029-5

Larssen, T., Høgåsen, T., Cosby, B.J. 2007. Impact of time series data on calibration and prediction uncertainty for a deterministic hydrogeochemical model, Ecol Model. 207: 22-33

Leavesley., G.H. 1994. Modelling the effects of climate change on water resources – a review. Climatic Change. 28: 159 - 173

Lenderink, G., Buishand, A., van Deursen, W. 2007. Estimates of future discharges of the river Rhine using two scenarios methodologies: direct versus delta approach, Hydrol Earth Syst Sci. 11 (4): 1145 - 1159

Lepistö, A., Etheridge, J.R., Granlund, K., Kotamäki,N., Maulve, O., Rankinen, K., and Varjopuro, R. 2013. Strategies to mitigate the impacts of climate change on European Freshwater Ecosystems. Report on the biophysical catchment-scale modelling of Yläneenjoki-Pyhäjärvi demonstration site. REFRESH EU Seventh Framework Programme

[LSEMS] Lake Simcoe Environmental Management Strategy. 1995. Lake Simcoe: Our Waters, Our Heritage. Lake Simcoe environmental management strategy implementation program summary of phase I progress and recommendations for phase II. In: Heathcote, I. (Ed.), Published by Lake Simcoe Region Conservation Authority, Ministry of Natural Resources, Ministry of Environment and Energy and Ministry of Agriculture, Food and Rural Affairs

[LSEMS] Lake Simcoe Environmental Management Strategy. 2010. Flow and nutrient monitoring data supplied by Lake Simcoe Region Conservation Authority. Available to download from http://www.lsrca.on.ca/programs/watershed_monitoring/system_intro.php

[LSRCA] Lake Simcoe Region Conservation Authority. 2006. Septic Tank Numbers. CANWET GIS Data Model.

[LSRCA] Lake Simcoe Region Conservation Authority. 2009. Report on phosphorus loads to Lake Simcoe 2004–2007. Lake Simcoe Region Conservation Authority, Ontario.

[LSRCA] Lakes Simcoe Environmental Management Strategy. 2010. Routine Monitoring Data Strategy O'Connor, E. Personal Communication

[LSRCA] Lake Simcoe Environmental Management Strategy. 2012a. Beaver River Subwatershed Plan. Durham Region.

[LSRCA] Lake Simcoe Region Conservation Authority. 2012b. Pefferlaw River Subwatershed Plan. Durham Region.

[LSRCA] Lake Simcoe Region Conservation Authority. 2012c. Holland River Subwatershed Plan. Durham Region

Maskey, S., Guinot, V,. Prices, R.K. 2004. Treatment of precipitation uncertainty in rainfall-runoff modelling: a fuzzy set approach, Adv Water Resour. 27: 889 - 898

McCollum, R.E. 1991. Buildup and decline in soil phosphorus: 30-year trends on a Typic Umprabuult, Agron J. 83: 77 – 85.

McDowell, R.W., Sharpley, A.N., Condron, L.M., Haygarth, P.M. and Brookes, P.C. 2001. Processes controlling soil phosphorus release to runoff and implications for agricultural management, Nut Cycl Agroecosys. 59: 269 – 284

McLauchlan, K. 2006. The nature and longevity of agricultural impacts on soil carbon and nutrients: a review, Ecosystems. 9: 1364 - 1382

McSweeney, C. and Jones, R. 2010. Selecting members of the 'QUMP' perturbed-physics ensemble for use with PRECIS. Met Office Hadley Centre

Meehl, G.A., and Stocker, T.F. 2007. Global climate projections in Climate Change: The Physical Science basis – Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Solomon, S., Quin, D., Manning, M., Chen, Z.., Marquis, M., Averyt, K.B., Tignor, M., and Miller, H.L. (Eds) Cambridge University Press, New York.

Miles, J. 2012. The relationship between land use and forms of phosphorus in the Beaver River watershed of Lake Simcoe, Ontario. MSc Thesis in Environmental and Life Sciences Graduate Program, Trent University

Nicholls, K.H., 1995. Some Recent water quality trends in Lake Simcoe, Ontario: implications for basin planning and limnological research, Can Water Resour J. 20 (4), 213–226.

Nicholls, K.H., and MacCrimmon, H.R. 1975. Nutrient Loading to Cook Bay of Lake Simcoe from the Holland River Watershed, Int Revue ges Hydrobiol. 60 (2) 159 – 193

Olding, A.B., Wicklund, R.E. and Richards, N.R. 1950. Soil Survey of Ontario County. Report No. 23 of the Ontario Soil Survey. Experimental Farms Service, Canada Department of Agriculture and the Ontario Agricultural College.

[OMAFRA] Ontario Ministry of Agriculture, Food and Rural Affairs. 2009. Agronomy Guide for Field Crops. [cited January 3rd, 2014] Available from http://www.omafra.gov.on.ca/english/crops/pub811/1toc.htm

Oni, S.K., Futter, M.N. and Dillon, P.J. 2011. Landscape-scale control of carbon budget of Lake Simcoe: A process based modelling approach, J Great Lake Res. 37: 160-165

Oni, S.K., Futter, M.N., Molot, L.A. and Dillon, P.J. 2014. Adjacent catchments with similar patterns of land use and climate have markedly different dissolved organic carbon concentration and runoff dynamics, Hydrol Process. 28: 1436 – 1449.

Ontario Ministry of Natural Resources. 2007. Ecological Land Classification of Ontario.

Oreskes, N., Shrader-Frechette, K., Belitz, K. 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263 (5147): 641-646

Paterson, A.M., Dillon, P.J., Hutchinson, N.J., Futter, M.N., Clark, B.J., Mills, R.B., Reid, R.A., and Scheider, W.A. 2006. A review of the components, coefficients and technical assumptions of Ontario’s Lakeshore Capacity Model, Lake Reserv Manage. 22 (1): 7-18

Peltouvouri, T. 2006. Phosphorus in agricultural soils of Finland - characterization of reserves and retention in mineral soil profiles. Helsinki: University of Helsinki, Pro Terra No 26. Academic dissertation.

[PGWMN] Provincial Groundwater Monitoring Network Program. 2012. Groundwater level data, groundwater chemistry data, and precipitation data. Ministry of Environment. Available from the world wide web: https://www.javacoeapp.lrc.gov.on.ca/geonetwork/srv/en/metadata.show?id=13677

[PWQMN] Provincial Water Quality Monitoring Network. 2009. Ministry of Environment.

Ramwekellan, J., Gharabaghi, B., Winter, J.G. 2009. Application of weather radar in estimation of bulk atmospheric deposition of total phosphorus over Lake Simcoe, Can Water Resour J. 34:37 – 60.

Regonda, S., Rajagopalan, B., Clark, M. and Pitlick, J. 2005. Seasonal cycle shifts in hydroclimatology over the western United States, J Clim. 18: 372–384. doi:10.1175/JCLI-3272.1.

Saelthun, N. 1995. Nordic HBV Model. Norwegian Water Resources and Energy. Administration

Schlotzhauer, S.M. and Price, J.S. 1999. Soil water flow dynamics in a managed cutover peat field, Quebec: Field and laboratory investigations, Water Resources Research 35 (12): 3675 – 3686

Scott, L.D. and Winter, J.G. 2006. Annual water balances, total phosphorus budgets and total nitrogen and chloride loads for Lake Simcoe (1998-2004). Lake Simcoe Environmental Management Strategy Implementation Phase III. Technical Report No. Imp. A.6.

Seibert, J. 1997. Esimation of parameter uncertainty in the HBV model. Nord. Hydrol. 28 (4/5). 246 - 262

Sharpley, A.N., Kleinman, P. and McDowell, R. 2001. Innovative management of agricultural phosphorus to protect soil and water resources, Commun Soil Sci Plan. 32 (7-8): 1071 – 1100

Smith, V.H. 2003. Eutrophication of freshwater and coastal marine ecosystems. *Environmental Science and Pollution Research.* 10 (2): 126 – 139

Smith, G.J. 2010. Deriving spatial patterns of severe rainfall in Southern Ontario from rain gauge and radar data. *A thesis presented to the University of Waterloo.* Waterloo, Ontario.

Starrfelt, S., and Kaste, O. 2014. Bayesian uncertainty assessment of a semi-distributed integrated catchment model of phosphorus transport, Envir Sci Process Imp. In Press

Statistics Canada. 2011. Farm and Operator Data. Census of Agriculture

Stephens, S.L.S. 2007. Optimizing agricultural and urban pollution remediation measures using watershed modeling: Review, calibration, validation and applications of the CANWET model in the Lake Simcoe watershed, Ontario, Canada. Trent University MS thesis 2007.

Teutschbein, C., and Seibert, J. 2012. Bias correction of regional climate model simulations for hydrological climate-change impact studies: review and evaluation of different models. *Journal of Hydrology* 456 – 457: 12 - 29

Thomas, R.L. and Sevean, G. 1985. Leaching of phosphorus from the organic soils of the Holland Marsh. A report to the Ministry of Environment.

Uhlenbrook, S., Seibert, J., Leibundgut, C., Rodhe, A. 1999. Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure. *Hydrol. Sci.* J. 44 (5): 779-797

[UKCP] UK Climate Projections. 2010. Probability in UKCP09 [cited October 25th 2014] Available from http://ukclimateprojections.metoffice.gov.uk/21680

[UKCP] UK Climate Projections. 2012. Perturbed Physics Ensembles. *Coupled Model Intercomparison Project.* . [cited January 3rd, 2014] Available from http://www.ukclimateprojections.metoffice.gov.uk/23251

Väänänen, R. 2008. Phosphorus retention in forest soils and the functioning of buffer zones used in forestry. *Dissertationes Forestales* 60. Department of Forest Ecology, University of Helsinki. P 42.

van den Hurk, B., Hirshci, M., Shcär, C., Lenderink, G., van Meijgaard, E., van Ulden, A., Rockel, B., Hagemann, S., Graham, P., Kjellström, E. and Jones, R. 2004. Soil control on runoff response to climate change in regional climate model simulations. J Climate. 18: 3536 – 3551

van Roosmalen, L., Christensen, B.S.B., Sonnenborg, T.O. 2006. Regional differences in climate change impacts on groundwater and stream discharge in Denmark, Vadose Zone J. 6: 554 - 571

Wade, A.J., Hornberger, G.M., Whitehead, P.G., Jarvie, H.P. and Flynn, N. 2001. On modelling the mechanisms that control in-stream phosphorus, macrophyte, and epiphyte dynamics: an assessmemt of a new model using general sensitivity analysis, Water Resour Res. 37 (11) 2777-2792

Wade, A.J., Whitehead, P.G. and Butterfield, D., 2002a. The Integrated Catchments model of Phosphorus dynamics (INCA-P), a new approach for multiple source assessment in heterogeneous river systems: model structure and equations, Hydrol Earth Syst Sci. 6, 583–606.

Wade, A.J., Whitehead, P.G. and O'Shea, L.C.M. 2002b. The prediction and management of aquatic nitrogen pollution across Europe: an introduction to the Integrated Nitrogen in European Catchments project (INCA), Hydrol Earth Syst Sci. 6, 299–313.

Wade, A., Butterfield, D., Lawrence, D.S., Bärlund, I., Durand, P., Lazar, A. and Kaste, O. 2007a. The integrated catchment model of phosphorus dynamics (INCA-P), a new structure to simulate particulate and soluble phosphorus transport in European catchments. Diffuse Phosphorus Loss: Risk Assessment, Mitigation options and Ecological Effects in River Basins. 5th Int

Phosphorus Workshop (IPW5), 3–7 September 2007 in Silkeborg, Denmark. Aarhus, Aarhus Universite: In: Heckrath,

Wade, A.J., Butterfieldm D., Griffiths, T. and Whitehead, P.G. 2007b. Eutrophication control in river-systems: an application to the River Lugg, Hydrol Earth Syst Sci. 11(1):584–600.

Westerman, P.W and Overcash, M.R. 1980. Short-term attenuation of runoff pollution potential for land-applied swine and poultry manure. Livestock waste- a renewable resource. Proceedings of the 4th International Symposium on Livestock Wastes, Am Soc Agric Eng. Sto Joseph, MI.

Whitehead, P.G., Wilby, R.L., Battarbee, R., Kernan, M. and Wade, A. 2009. A review of the potential impacts of climate change on surface water quality, Hydrolog Sci J. **54**, 101-123.

Whitehead, P.G., Jin, L., Baulch, H.M., Butterfield, D., Oni, S.K., Dillon, P.J., Futter, M., Wade, A.J., North, R., O'Connor, E.M. and Jarvie, H.P. 2011. Modelling phosphorus dynamics in multi-branch river systems: a study of the Black River, Lake Simcoe, Ontario, Canada, Sci Total Environ. 412–413, 315–323.

Whitehead., P.G., Jin, L., Crossman, J., Comber, S., Johnes, P.J., Daldorph, P., Flynn, N., Collins, A.L., Butterfield, D., Mistry, R., Bardon, R., Pope, L. and Willows, R. 2014. Distributed and dynamic modelling of hydrology, phosphorus and ecology in the Hampshire Avon and Blashford Lakes: evaluating alternative strategies to meet WFD standards, Sci Total Environ. 481: 157 - 166

Wilby, R.L. and Harris, I. 2006. A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames UK, Water Resour Res. 42: 1-10. Doi: 10.1029/2005WR004065

Winter, J.G., Dillon, P.J., Futter, M.N., Nicholls, K.H., Wolfgang, A.S. and Scott, L.D. 2002. Total phosphorus budgets and nitrogen loads: Lake Simcoe, Ontario (1990-1998), J Great Lakes Res. 28 (3): 301-314

Winter, J.G., Eimers, M.C., Dillon, P.J., Scott, L.D., Scheider, W.A. and Campbell, C.W. 2007. Phosphorus inputs to Lake Simcoe from 1990 to 2003: Declines in tributary loads and observations on Lake Water Quality, J Great Lakes Res. 333:381 - 396

XCG Consultants. 2010. Review of Phosphorus Removal at Municipal Sewage Treatment Plants Discharging to the Lake Simcoe Watershed. Prepared for Water Environment Association of

Ontario Lake Simcoe Clean-Up Fund. Environment Canada.

8. Tables

Land Type		% Land Cover			
		Holland	Pefferlaw	Beaver	Whites
Agriculture	Intensive	34.9	23.4	28.1	20.4
	Non Intensive	17.8	25.3	37.0	45.7
Urban		19.6	10.6	5.1	1.7
Wetland		10.8	18.2	19.3	21.2
Forest		16.9	22.5	10.5	11.0

Table 1: Landuse cover of the four study catchments within the Lake Simcoe Catchment *reported to 1 decimal place (1 d.p)*

Parameter	Data description	Study Site				Data Source
		Holland	Pefferlaw	Beaver	Whites	
Catchment characteristics	Catchment area (km2)	613.8	444.4	325.3	85.0	Modelled using a 2m vertical resolution DEM (Global Land Cover Facility; 2002)
	Number of sub catchments	39	41	30	23	
	Quaternary Geology (% clay composition)	8.3	12.6	27.5	41.9	Catchment average calculations from GIS maps derived from Agriculture and Agri-Food Canada (2010)
Hydrological characteristics	Temperature (oC) daily data	7.4	7.3	7.3	7.3	Measurements from local Environment Canada meteorological stations
	Precipitation daily data	2.5	2.5	2.5	2.0	
	SMD (mm)	42.8	48.0	47.2	22.1	Derived from HBV model
	HER (mm)	0.5	0.9	1.1	0.7	
	Soil water residence time (days)	3.0	2.7	2.4	1.4	Calculations from flow hydrographs derived from field monitoring data (Dave Woods, *personal communication;* LSRCA 2010; LSEMS 2010; PWQMN, 2009)
	Saturation excess threshold (m^3/s)	0.6	3E-2	0.5	0.5	
	Flow-velocity relationship (flow a and b parameters)	a: 6.0E-2 b: 0.70	a: 5.0E-2 b: 0.20	a: 8.0E-2 b: 0.67	a: 6.0E-2 b: 0.7	Derived from monitoring data (Dave Woods, *personal communication;* LSRCA 2010; LSEMS 2010; PWQMN, 2009)
P budget in non-intensive Agriculture	Fertiliser inputs from grazing animals and applications to crops (kg/ha/yr)	33.2	32.4	30.0	25.8	Calculations using data from Statistics Canada (2011); OMAFRA (2009); Bangay (1976) using methods of Wade et al (2007b)
	Septic tank inputs (kg/ha/year)	1.2	0.8	0.2	0.2	Calculations based on data from Statistics Canada (2011) and Scott et al (2006);using methods of Paterson et al (2006) and Stephens (2007)
	Maximum Plant Uptake (kg/ha/year)	100	100	100	100	Based on previous INCA applications to the Simcoe region by Jin et al (2013)
P budget in intensive Agriculture	Fertiliser inputs from grazing animals and applications to crops (kg/ha/yr)	21.2	24.0	32.5	32.4	Calculations using data from Statistics Canada (2011); OMAFRA (2009); Bangay (1976) using methods of Wade et al (2007b)
	Maximum Plant Uptake (kg/ha/year)	100	100	100	100	Based on previous INCA applications to the Simcoe region by Jin et al (2013)
P inputs to whole catchments	Atmospheric Deposition: regional values (kg/ha/yr)	0.21	0.21	0.21	0.21	Regional monitoring data (Ramwekellan et al., 2009)
	Groundwater TDP concentration (mg/l)	7.0E-3	6.0E-3	1.0E-1	7.0E-7	Based on information from the Provincial Groundwater Monitoring Network (PGWMN, 2014)
Sewage (STW) P inputs to river reaches	Number of STWs	3	1	2	N/A	XCG consultants Ltd (2010) and KMK consultants (2004)
	Average Inputs (kg P/year)	103.6	102.9	67.2	N/A	

Table 2: Details of key parameters and data sources used in model calibration (1 d.p)

QUMP member	Climate sensitivity
Q5	2.2
Q9	2.2
Q1	2.6
Q15	**2.9**
Q16	3.0
Q2	3.1
Q14	3.1
Q13	**3.2**
Q11	3.2
Q7	3.3
Q	3.4
Q0	**3.5**
Q6	3.8
Q3	**3.8**
Q10	**3.9**
Q4	4.6
Q8	4.9
Q12	4.9

Table 3: Quantifying Uncertainty in Model Predictions (QUMP) members listed in order of climate sensitivity under a doubling of CO_2 concentrations (from Collins et al., 2006). Bold members indicate those selected for use in the study.

Catchment	Maximum Model coefficient					Model coefficient at downstream extent				
	Flow		TP Concentration		TP loads*	Flow		TP Concentration		TP loads*
	R^2	Model Error (MAE) (%)	R^2	MAE (%)	R2	R^2	MAE (%)	R^2	MAE (%)	R^2
Holland	0.95	-24.78	0.72	-4.36	0.74	0.95	-24.78	@0.52	@-20.12	@0.60
Pefferlaw	0.91	+14.06	0.37	-23.4	0.71	0.91	+46.28	0.34	-23.41	0.61
Beaver	0.98	-15.30	0.79	+4.40	0.82	0.85	33.50	0.05	-58.37	0.72
Whites	0.92	+3.03	0.47	-26.34	0.94	0.92	+3.03	0.29	-26.34	0.77

Table 4: Model performance statistics presented for locations of highest calibration accuracy and at downstream extent (for which corresponding observed flow, TP and loads data were available). Results are reported to 2.d.p. For individual performance statistics of each tributary, see Figure 6 [a Due to lack of observed data available for model performance assessment in the convergence zone of the East and West branch of the Holland, reported outflow performance statistics are an average

(mean) of statistics from the two individual branches. *Due to the different temporal resolution of sampling vs modelled data “observed loads” are not directly comparable with modelled loads. Both are calculated using monthly means of flow and total phosphorus data, however, although degree of load variance explained by the model can be determined (R^2), a direct quantification of over or under-estimation (model error) cannot be calculated.]

	Source	Coefficients (kg P/km^2/yr)				Source (% contribution)			
		East/West Holland	Pefferlaw	Beaver	Whites	East/West Holland	Pefferlaw	Beaver	Whites
In-stream	STW	0.8	0.3	0.5	0.00	5.5	3.7	13.3	0.00
Soils	Urban	9.9	4.8	2.4	2.1	14.0	6.1	3.1	0.4
	Intensive Agriculture	16.1	8.3	3.7	8.7	40.8	23.0	26.3	21.2
	Non Intensive Agriculture	16.0	13.8	4.2	8.9	20.6	41.3	38.9	48.7
	Wetlands	14.4	5.1	2.5	8.1	11.3	11.0	12.1	20.4
	Forest	6.4	5.6	2.4	7.1	7.8	14.9	6.3	9.3
	Average soils export	**13.1**	**8.1**	**3.4**	**8.5**	**100**	**100**	**100**	**100**

Table 5: Export coefficients (kg/km2/year) of total phosphorus from diffuse and point sources, and contributions of each source as a percentage of total phosphorus export *(1 d.p.).*

Date	Catchment	Average Uncertainty				
		Temperature (oC)	Precipitation (%)	Flow (%)	TP (%)	Loads (%)
2030	Holland	0.9	19.6	34.0	5.9	39.3
	Pefferlaw	0.9	16.2	23.3	7.8	27.9
	Beaver	0.9	16.9	29.0	11.3	36.6
	Whites	0.9	19.7	43.2	20.8	58.9
	Range	3.0E-2	3.5	19.9	15.0	30.9
2070	Holland	1.9	23.7	37.3	7.1	43.0
	Pefferlaw	1.9	23.1	27.6	11.1	33.4
	Beaver	1.9	23.2	35.1	11.9	38.7
	Whites	2.0	22.60	40.1	23.6	53.8
	Range	0.1	1.1	12.5	16.5	20.4

Table 6: A measure of average uncertainty of temperature, precipitation, flow, TP concentrations and loads across the Perturbed Physics Ensemble, where “average uncertainty” is calculated as the mean value of uncertainty across all probability levels within the cumulative distribution function *(1 d.p)*

Catchment	Change in HER per mm change in precipitation			% surface runoff		% soil matrix flow		Soil water residence times (days)	Clay Content (%)
	2030	2070	Average	2030	2070	2030	2070		
Holland	4.9E-1	5.5E-1	5.2E-1	6.1	5.7	93.9	94.3	3.0	8.3
Pefferlaw	8.0E-1	9.3E-1	8.7E-1	12.3	11.2	87.7	88.8	2.7	12.6
Beaver	8.9E-1	9.9E-1	9.4E-1	16.1	31.8	83.9	68.2	2.4	27.5
Whites	8.6E-1	1.0	9.4E-1	22.6	31.6	77.4	68.4	1.4	41.9

Table 7: Calculation of quantity of hydrologically effective rainfall (HER) (mm) generated per mm of precipitation *(1.d.p.),* and comparison with catchment characteristics (hydrological and quaternary geology)

		Mean annual % change in TP loads					
		Q0	Q3	Q10	Q13	Q15	**EA**
2030	Holland	22.6	11.7	-8.0	10.8	3.7	**8.2**
	Pefferlaw	13.5	10.5	-2.9	-3.6	-6.6	**2.2**
	Beaver	4.5	0.2	-12.8	-7.6	-16.4	**-6.4**
	Whites	13.5	2.4	-9.0	-4.2	-25.1	**-4.5**
	Range						**14.6**
2070	Holland	29.6	17.6	-2.0	20.9	7.3	**14.7**
	Pefferlaw	13.2	9.1	-14.2	-5.7	-7.5	**-1.0**
	Beaver	0.3	-7.2	-26.3	-16.5	-23.8	**-14.7**
	Whites	7.0	-3.2	-30.5	-14.8	-24.1	**-13.1**
	Range						**29.4**

Table 8: Annual change in total phosphorus loads between baseline and future periods, for all ensemble members of the perturbed physics ensemble. Includes ensemble average (EA) change *(1 d.p)*

Catchment	Season	Projected change in TP (ug/l) per mm of change in precipitation	Proportion of annual change (%)	Clay content (%)
Holland	spring	5.9	6.7	
	summer	24.4	27.6	
	autumn	54.0	**61.1**	
	winter	4.1	4.6	
	Annual Average	22.1	100	8.3
Pefferlaw	spring	5.0	11.7	
	summer	23.8	**56.1**	
	autumn	4.0	9.5	
	winter	9.7	22.7	
	Annual Average	10.6	100	12.6
Beaver	**Spring**	201.0	**86.2**	
	Summer	4.9	2.1	
	Autumn	13.5	5.8	
	Winter	13.8	5.9	
	Annual average	58.3	100	27.5
Whites	**spring**	28.3	5.0	
	summer	5.5	1.0	
	autumn	18.2	3.2	
	winter	518.8	**90.9**	
	Annual Average	142.7	100	41.9

Table 9: Estimation of average catchment sensitivity of water quality to uncertainty in precipitation: projected change in total phosphorus (mg/l) per mm change in precipitation averaged over 2030’s and 2070’s *(1 d.p)*

Figure 1: Map of 4 study catchments within the Simcoe Basin, Southern Ontario, demonstrating PRECIS regional climate model resolution, available meteorological stations, and catchment boundaries (RCM = Regional Climate Model; PRECIS = Providing Regional Climates for Impacts Studies model)

Figure 2: Comparison of monthly temperature data from 5 ensemble members (Q0-15), regionally downscaled using PRECIS, with observed meteorological station data over 30 year baseline period (1968 – 1997)

Figure 3: Comparison of monthly precipitation data from 5 ensemble members (Q0-15), regionally downscaled using PRECIS, with observed meteorological station data over 30 year baseline period (1968 – 1997)

Figure 4: Calculations of A) monthly change in temperature and B) projected uncertainty from cumulative distribution functions of perturbed physics ensemble temperature data of the Holland River in 2030

Figure 5: Example of INCA model output. Calibration results from Holland River comparing modelled and observed data of a) flow at East Branch outflow b) total phosphorus (TP) concentrations at East Branch outflow c) TP concentration at West branch outflow

Figure 6: Spatial variability in model accuracy of total phosphorus (TP) concentrations throughout INCA models calibrated across multiple reaches, illustrating relative model error in A) Holland, B) Pefferlaw, C) Beaver and D) Whites catchments. Data represents average TP concentrations over the calibration period.

Figure 7: Analysis of INCA model fit to event data at outflow of A) Holland, B)Pefferlaw, C) Beaver and D) Whites catchments. Rising and recession limb at analysed separately for R^2 and model average error (MAE) to determine model performance under precipitation events. All events were analysed during the same four day period of intense precipitation in November 2011. Although event curves can be used to confirm model representation of reality, it should be noted they should not be used to compare response times between catchments.

Figure 8: Analysis of model response to observed spring melt in A) Holland, B) Pefferlaw C) Beaver and D) Whites catchments. Results analysed during spring melt of 2010.

Figure 9: Between-catchment comparison of the CDFs from the perturbed physics ensemble of A) monthly change in temperature by 2030, B) monthly change in temperature by 2070, C) monthly change in precipitation by 2030 D) monthly change in precipitation by 2070 (CDF: cumulative distribution function)

Figure 10: Between-catchment comparison of the CDFs from the perturbed physics ensemble of A) monthly change in flow by 2030, B) monthly change in flow by 2070, C) monthly change in total phosphorus (TP) concentration by 2030, D) change in TP concentration by 2070 (CDF: cumulative distribution function)

Figure 11: Between-catchment comparison of the CDFs from the perturbed physics ensemble of A) change in total phosphorus loads (TP) by 2030, B) change in TP loads by 2070 (CDF: cumulative distribution function

Figure 12: Boxplot of seasonal changes in total phosphorus concentrations across all perturbed physics ensemble members (one box per ensemble member) in A) Holland B) Pefferlaw C) Beaver and D) Whites