Anonymous Referee #3

The reviewer provides complementary comments on the quality of the paper but also indicates "…the authors test design and discussions are not adequate to derive the intended conclusions. The model configuration, calibration process are not adequately reported. Uncertainty analysis is missing. My general comment is that the study can be accepted if the authors are able to address the following shortcomings in sufficient detail and only after a major revision."

- In our revisions we have attempted to address the above issues raised by the reviewer. Additional text has been added to the Results and Discussion sections so that the reader can better understand how we reached our conclusions. We have also provided an extended text on the model configuration and calibration process, and have provided further details about the uncertainty analysis. Further responses are given in comments (below) in which we demonstrate specific changes made to the paper.

Specific comments:

a) Abstract

1. Page 4316, line 10, "comparison of simulated daily mean discharge… allowed the error in the model prediction to be quantified". The authors failed to properly address the claim they raised here, in the main body of paper.

- Discharge has been removed from the comparison. We were not able to do a comparison of observed high–frequency, event–based discharge measurements (2010–2012) against modelled daily mean simulations of discharge because the observed measurements at the FRI stream–gauge for the period 2010-2012 were not available. In July 2010, the gauge was repositioned 720 m downstream to the State Highway 30 (SH 30) bridge (Page 4320, lines 20–21).

2. (1) The authors suggested hiring higher frequencies of observation in order to overcome the base and quick flow dependent regimes limitations in current model. (Page 4316, line 15). (2) Please explain how this improve the model performance? (3) Do you also consider sub–daily simulations? Please clarify that in the proper section in the main text.

- (i) The statement has been added in the text as follows: "We did not use the high–frequency observations to calibrate the model, because of the limited number of high–frequency (1–2 h) samples (nine events for SS and 14 events for TP and TN in 2010–

2012). The use of the high–frequency observations for model validation allowed to examine how the model performed during short (1–3 day) high flow periods".

(ii) To describe how the model reproduced the data derived from the high–frequency observations, we have showed simulated results as follows: "Monthly instantaneous TN concentrations used for model calibration and validation were generally not reproduced well in simulations ($R^2 < 0.1$ and NSE < 0). The model showed satisfactory performance ($R^2$and NSE both ~0.5) in reproducing daily mean discharge–weighted TN concentrations derived from high–frequency measurements (1–2 h) taken over 14 storm events of duration 24–73 h". Therefore, we have stated as follows: "To address this, we recommend that high–frequency, event–based monitoring data are used to support calibration and validation".

(iii) In relation to sub–daily simulations, please see the response to *Reviewer #1, comment #5*: We did not consider sub–daily simulations. The version of the SWAT model used in this study (SWAT2009_rev488) runs on a daily time step. This has been added at the beginning of Section Model configuration as follows: "The SWAT model version used (SWAT2009_rev488) runs on a daily time step". We provide additional reasoning for not using sub–daily time steps, as mentioned in Table 1 as follows: "measurements for important meteorological forcing variables (e.g., temperature, relative humidity and solar radiation) were available only at daily resolution".

3. Abstract, page 4316, line 17, again you are thronging an idea that your study has implications in identifying uncertainties but you are very inexact in explaining how?

- We have revised the text in the Abstract to better explain the identification of uncertainties, as follows: "This study has important implications for identifying uncertainties in parameter sensitivity and performance of hydrological models applied to catchments with large fluctuations in stream flow, and in cases where models are used to examine scenarios that involve substantial changes to the existing flow regime".

4. Please be very specific of the outcome of this study in your abstract. Make 2-3 bullet points of what you achieved during this study.

- Please see our response to *Reviewer #1, comment #2* and *Reviewer #2, comment #2*. We have not put bullet points in the Abstract as this would not conform to the usual format of

an abstract. However, we have included additional text to capture the main findings of the study as follows: "Monthly instantaneous TP and TN concentrations were generally not reproduced well (24% bias for TP, 27% bias for TN, and $R^2 < 0.1$, NSE < 0 for both TP and TN), in contrast to SS concentrations (< 1% bias; R^2 and NSE both > 0.75) during model validation. Comparison of simulated daily mean SS, TP and TN concentrations with daily mean discharge–weighted high–frequency measurements during storm events indicated that model predictions during the high rainfall period considerably underestimated concentrations of SS (44% bias) and TP (70% bias), while TN concentrations were comparable (< 1% bias; R^2 and NSE both ~0.5). Several SWAT parameters were found to have different sensitivities between base flow and quick flow. Parameters relating to main channel processes were more sensitive for the base flow estimates, while those relating to overland processes were more sensitive for the quick flow estimates".

b) Introduction

5. Page 4318, line 10, "They found that the logarithmic form of the Nash-Sutcliffe efficiency (NSE) value provided more information on the sensitivity of model performance for simulations of discharge during storm events, while the relative form of NSE was better for base flow periods." this is not what Krause et al (2005) had been reported. In their paper they clearly stated that: "To reduce the problem of the squared differences and the resulting sensitivity to extreme values the Nash-Sutcliffe efficiency E is often calculated with logarithmic values of O and P. Through the logarithmic transformation of the runoff values the peaks are flattened and the low flows are kept more or less at the same level. As a result the influence of the low flow values is increased in comparison to the flood peaks resulting in an increase in sensitivity of ln E to systematic model over- or underprediction". Beside they used natural logarithm and not log 10. I also couldn't find the justification for the threshold number "0.1". Please clarify this.

- Krause et al. (2005) stated in section 2.5 that "The logarithmic form of E [Nash–Sutcliffe efficiency] is widely used to overcome the oversensitivity to extreme values", and in section 2.6 "it can be expected that the relative forms are more sensitive on systematic over– or underprediction, in particular during low flow conditions". We took this latter statement to mean that: "the logarithmic form of the Nash–Sutcliffe efficiency (NSE) value provided more information on the sensitivity of model performance for discharge

simulations during storm events, while the relative form of NSE was better for base flow periods" (see page 4318 lines 11–14). Therefore the natural logarithm was used by Krause et al. (2005) and therefore the standard deviation (STD) of the ln–transformed NSE were used to indicate parameter sensitivity for the two flow regimes.

We have clarified the justification for the threshold in the paper text as follows: "The threshold value of "0.2" was chosen in this study, based on the median value derived from the calculations of the STD of ln–transformed NSE".

c) Parameter calibration (I would call it model calibration!)

6. Page 4321, line 9. Latin hypercube method is a sampling method that insures the samples cover the entire parameter space and that the optimum solution is not a local minimum. LH is not quantifying uncertainties… please correct for that.

- The heading has been changed to state model calibration and validation, as suggested.

With regard to the Latin hypercube sampling method, we have not altered the sentence on Page 4321, lines 10–11 but have added text that the reviewer suggested as follows: "Latin hypercube sampling (LHS) is a method that generates a sample of plausible parameter values from a multidimensional distribution and ensures that samples cover the entire parameter space, therefore ensuring that the optimum solution is not a local minimum (Marino et al., 2008)".

7. The calibration process is very vague to a non-swat user. Please give adequate information on calibration steps. You jumped from LH to R factor and P factor…describe your calibration procedure in short but sufficiently. Your calibration set up is unclear.

(1) Did you calibrate discharge and sediment and nitrate all together or one after the other?

(2) How did you select your parameters at first place?

(3) Did you perform some sensitivity analysis prior to calibration?

(4) Page 4321, line 16, "produce narrower parameter range", how?

(5) How many simulations you had? How many iterations?

(6) Page 4321, line 17, "optimal value.." how do you know? ref?

(7) What are the fitted value for the selected parameter after calibration (best parameter set)?

- We have altered the manuscript in the sub–section Model calibration and validation in response to the need to provide information on the calibration steps:

  (i) The sequence of calibration is described on Page 4322, lines 13–16 and the text has been better clarified by rearranging as follows: "Daily mean discharge was firstly calibrated based on daily mean values of 15–minute measurements. Water quality variables were then calibrated in the sequence: SS, TP and TN. Modelled mean daily concentrations were compared with concentrations measured during monthly grab sampling, with monthly measurements assumed equal to daily mean concentrations".

  (ii) We selected parameter values as follows (Page 4320, lines 26–27 and Page 4321, lines 1–2): "Values of SWAT parameters were assigned based on: i) measured data (e.g. some of the soil parameters; Table 1); ii) literature values from published studies of similar catchments (e.g. parameters for dominant land uses; Table 2); or iii) by calibration where parameters were not otherwise prescribed".

  (iii) Please see *Reviewer #2, Comment #6 (iv)*. Steps and equations used in the SUFI–2 procedure to analyse parameter sensitivity are outlined by Abbaspour et al., (2004). The procedure of sensitivity analysis has been briefly described in new text as follows: "The SUFI–2 procedure analyses relative sensitivities of parameters by randomly generating combinations of values for model parameters (Abbaspour et al., 2014). A sample size of 1000 was chosen for each iteration of LHS, resulting in 1000 combinations of parameters and 1000 simulations. Model performance was quantified for each simulation based on the Nash–Sutcliffe efficiency (NSE). An objective function was defined as a linear regression of a combination of parameter values generated by each LHS against the NSE value calculated from each simulation. Each compartment was not given weight to formulate the objective function because only one variable was specifically focused on at each time. A parameter sensitivity matrix was then computed based on the changes in the objective function after 1000 simulations. Parameter sensitivity was quantified based on the p value from a Student's t–test, which was used to compare the mean of simulated values with the mean value of measurements (Rice, 2006). A parameter was deemed sensitive by if $p \leqslant 0.05$ after 1000 simulations (one iteration). Numerous iterations of LHS were conducted. Values of p from numerous iterations were averaged for each parameter, and the frequency of iterations where a parameter was deemed sensitive was

summed. Rankings of relative sensitivities of parameters were developed based on how frequently the sensitive parameter was identified and the averaged value of *p* calculated from several iterations. The most sensitive parameter was determined based on the frequency that the parameter was deemed sensitive, and the smallest average *p*–value from all iterations".

A new table has also been added in the text to show the ranking of relative sensitivities of hydrological and water quality parameters derived from the SUFI–2 procedure. The text has been added in Method as follows: "A one–at a–time (OAT) routine proposed by Morris (1991) was applied to investigate how parameter sensitivity varied between the two flow regimes (base flow and quick flow), based on the ranking of relative sensitivities of parameters that were identified by randomly generating combinations of values for model parameters for each individual variable using the SUFI–2 procedure". The text has also been added in Results as follows: "Based on the ranking of relative sensitivities of hydrological and water quality parameters derived from the SUFI–2 procedure (see Table 7), the OAT sensitivity analysis undertaken separately for base flow and quick flow identified…".

Table 7 Rankings of relative sensitivities of parameters (from most to least) for variables (header row) of Q (discharge), SS (suspended sediment), MINP (mineral phosphorus), ORGN (organic nitrogen), NH_4–N (ammonium–nitrogen), and NO_3–N (nitrate–nitrogen). Relative sensitivities were identified by randomly generating combinations of values for model parameters and comparing modelled and measured data with a Student's t test ($p \leqslant 0.05$). Bold text denotes that a parameter was deemed sensitive relative to more than one simulated variable. Shaded text denotes that parameter deemed insensitive to any of the two flow components (base and quick flow; see Figure 7) using one–at a–time sensitivity analysis. Definitions and units for each parameter are shown in Table 3.

Q	**SS**	**MINP**	**ORGN**	**NH_4–N**	**NO_3–N**
SLSOIL	LAT_SED	CH_OPCO	**CH_ONCO**	**CH_ONCO**	NPERCO
CH_K2	CH_N2	BC4	BC3	BC1	**CDN**
HRU_SLP	SLSUBBSN	RS5	SOL_CBN(1)	**CDN**	ERORGN
LAT_TTIME	SPCON	ERORGP	RS4	RS3	**CMN**
SOL_AWC(1)	ESCO	PPERCO	**RCN**	**RCN**	**RCN**
RCHRG_DP	OV_N	RS2	N_UPDIS		**RSDCO**
GWQMN	**SLSOIL**	PHOSKD	USLE_P		
GW_REVAP	**LAT_TTIME**	GWSOLP	SDNCO		
GW_DELAY	**SOL_AWC(1)**	LAT_ORGP	SOL_NO3(1)		
CH_COV1	**EPCO**		**CMN**		
CH_COV2	**CANMX**		HLIFE_NGW		
EPCO	**CH_K2**		**RSDCO**		
SPEXP	**GW_DELAY**		USLE_K(1)		
CANMX	ALPHA_BF				
CH_N1	**GW_REVAP**				
PRF	**CH_COV1**				
SURLAG					

(iv) Steps and equations used in the SUFI–2 procedure to constrain parameter ranges are outlined by Abbaspour et al., (2004). The method to produce narrower parameter ranges has been briefly described in new text in the paper as follows: "A range was first defined for each parameter based on a synthesis of ranges from similar studies or from the SWAT default range. Parameter ranges were updated after each iteration based on the computation of upper and lower 95% confidence limits. The 95% confidence interval and the standard deviation of a parameter value were derived from the diagonal elements of the covariance matrix, which was calculated from the sensitivity matrix and the variance of the objective function. Steps and equations used in the SUFI–2 procedure to constrain parameter ranges are outlined by Abbaspour et al. (2004)".

(v) The number of simulations and iterations has been described in the Method as follows: "A sample size of 1000 was chosen for each iteration of LHS, resulting in 1000 combinations of parameters and 1000 simulations. Numerous iterations (each comprising 1000 samples) of LHS were conducted. The total numbers of iterations performed for each simulated variable (Q, SS, MINP, ORGN, NH_4–N and NO_3–N) reflected the numbers required to ensure that > 90% of measured data were bracketed by simulated output and the R–factor was close to one."

The relevant text has also been added in the Results as follows: "Numerous rounds (each comprising 1000 iterations) of LHS were conducted for each simulated variable until the performance criteria were satisfied. The total number of rounds of LHS for each simulated variable was as follows (number in parentheses): Q (7), SS (7), MINP (11), ORGN (10), NH_4–N (4) and NO_3–N (4)".

(vi) The process for derivation of the optimal parameter values has been described in new text as follows: "The 'optimal' parameter value was obtained when the Nash–Sutcliffe efficiency (NSE) criterion was satisfied (NSE > 0.5; Moriasi et al., 2007)".

(vii) The statement has been added in the text as follows: "The parameters that provided the best statistical outcomes (i.e, best match to observed data) are given in Table 3".

8. You referred to R–factor and P–factor but you didn't perform uncertainty analysis or at least you didn't report it! This indices are not used later on in the text! How wide is the uncertainty range? What are the possible explanation for that?

- SUFI–2 considers two criteria to constrain parameter ranges in each iteration (Abbaspour, 2014). One is the P–factor, the percentage of measured data bracketed by 95% prediction uncertainty (95PPU). Another is the R–factor, the average thickness of the 95PPU band divided by the standard deviation of measured data. The text added in the Methods reads: "Model uncertainty was evaluated by two criteria; R–factor and P–factor (see Section 2.3). They were used to constrain parameter ranges during the calibration using measured Q and loads of SS, MINP, ORGN, NH_4–N and NO_3–N in the SUFI–2 procedure". The values of the R–factor and P–factor were automatically updated in SUFI–2 during the auto–calibration and their final results have been reported in the text as follows: "Two criteria (R–factor and P–factor) were used to show model uncertainties for simulations of discharge and contaminant loads, with values as follows: Q (0.97, 0.43), SS (0.48, 0.19), MINP (2.64, 0.14), ORGN (0.47, 0.17), NH_4–N (1.16, 0.56) and NO_3–N (1.2, 0.29)".

  We compared the measured and simulated SS and TN concentrations using the auto–calibrated parameters. We used manual calibration based on the measured TP concentration. Additionally, we have also analysed model uncertainties by graphically showing the 95% confidence interval for measurements and the 95% prediction interval for model simulations of Q and SS, TP and TN concentrations. The text added in the Methods reads: "The R software was used to graphically show the 95% confidence and prediction intervals for measurement data (Neyman, 1937) and model prediction intervals (Seymour, 1993) for Q and concentrations of SS, TP and TN during the calibration period (2004–2008)". The text added in the Results reads: "Model uncertainties for simulations of Q and SS, TP and TN concentrations are shown in Fig. 6".

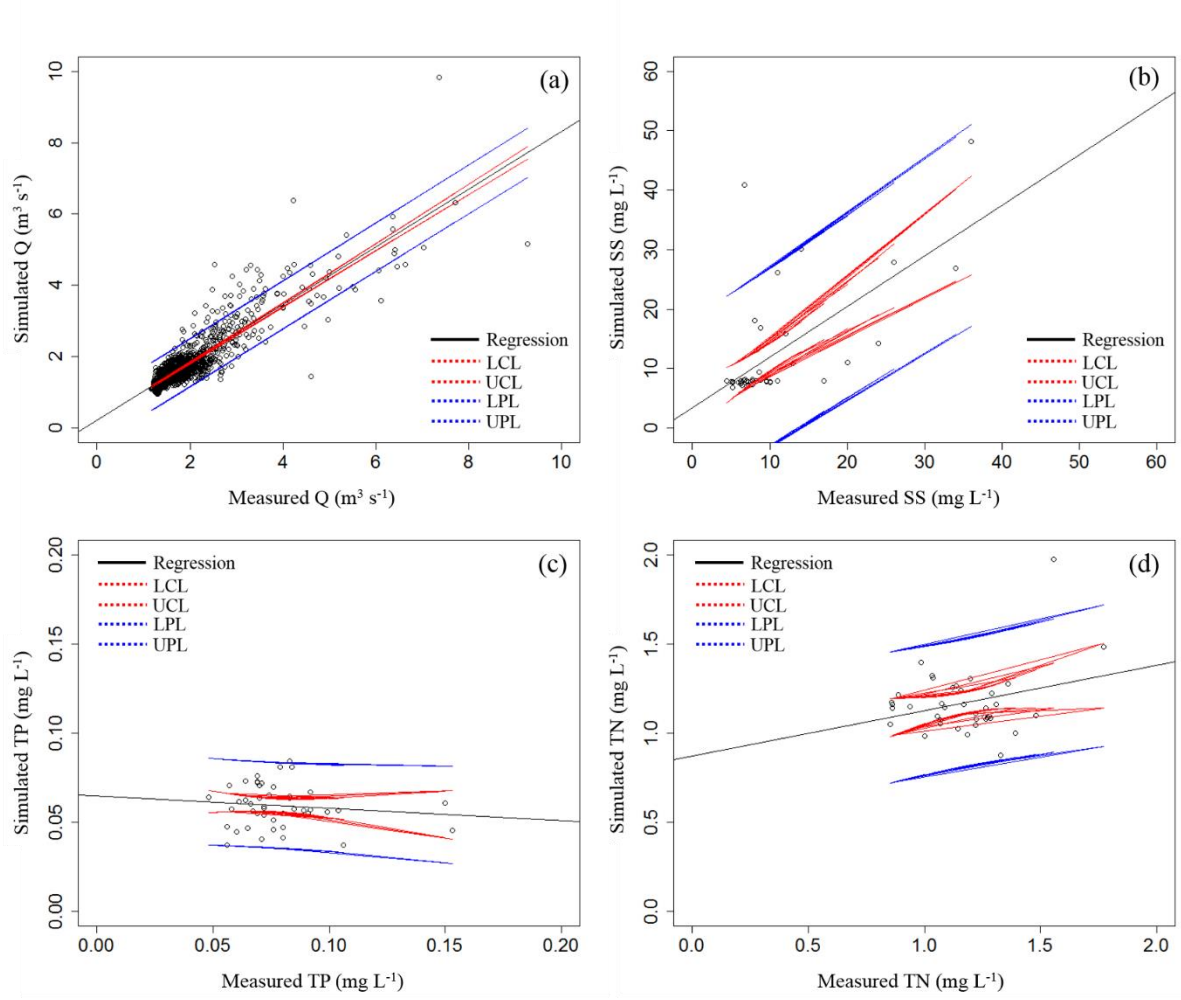

Figure 6. Regression of measured and simulated (a) discharge (Q), concentrations of (b) suspended sediment (SS), (c) total phosphorus (TP), and (d) total nitrogen (TN) including lower and upper 95% confidence limits (LCL and UCL) and lower and upper 95% prediction limits (LPL and UPL). Note that the "choppy" shape of confidence limits shown in figures b–d were resulted from the few data points (< 50) in the regressions of measured and simulated SS, TP and TN concentrations.

Explanation of model uncertainty has been added in the Discussion as follows: "Model uncertainty in this study may arise from four main factors: 1) model parameters; 2) forcing data; 3) in measurements used for evaluation of model fit, and; 4) model structure or algorithms (Lindenschmidt et al., 2007). The values of most parameters assigned for model calibration, although specific to different soil types (e.g. soil parameters), were lumped across land uses and slopes in this study. They integrated spatial and temporal variations, thus neglecting any variability throughout the study catchment. In terms of forcing data, the assumption of constant values of spring discharge

rate and nutrient concentrations may inadequately reflect the temporal variability and therefore increase model uncertainty, although this should contribute little to the model error term. Most water quality data used for model calibration comprised monthly instantaneous samples taken during base flow conditions. The use of those measurements for model calibration would likely lead to considerable underestimation of constituent concentrations (notably SS and TP) due to failure to account for short–term high flow events. Inadequate representation of groundwater processes in the model structure is another key factor that is likely to affect model uncertainty, particularly for nitrogen simulations. The analysis of model performance based on datasets separated into base flow and quick flow constituents enabled uncertainties in the structure of hydrological models to be identified, denoted by different model performance between these two flow constituents". Another discussion on Page 4329, lines 19–26 said: "Furthermore, the disparity in goodness–of–fit statistics between discharge (typically 'good' or 'very good') and nutrient variables (often 'unsatisfactory') highlights the potential for catchment models which inadequately represent contaminant cycling processes (manifest in unsatisfactory concentration estimates) to nevertheless produce satisfactorily load predictions (e.g., compare model performance statistics for prediction of nutrient concentrations in Table 5 with statistics for prediction of loads in Table 6). This highlights the potential for model uncertainty to be underestimated in studies which aim to predict the effects of scenarios associated with changes in contaminant cycling, such as increases in fertiliser application rates".

d) Model evaluation

9. Page 4322, line 1-10, (1) SWAT accounts for initial amount of Nitrate in shallow groundwater and the corresponding parameter is NO3 sh,o. (2) To the extent of my knowledge you can also set your background N in soil layers. (3) "Model general underestimation" is not a valid justification to add 0.44 mg N L^{-1} to your model simulation. You need either to adjust your input or have stronger argument for doing so.

- (i) In SWAT documentation, there is indication that the initial nitrate concentration in the shallow groundwater has been considered, but there is no command to input a value and run the model.

  (ii) It would not be appropriate to add this parameter value into the initial soil nitrogen as suggested, because the transport processes are different.

(iii) This argument has been added in the text as follows: "Over the period of the first five years of wastewater irrigation, nitrate concentrations in shallow groundwater draining the Waipa Stream sub–catchment were estimated to have increased by c. 0.44 mg L^{-1} (Paku, 2001). SWAT has no capability to dynamically adjust the groundwater concentration during a simulation run. Therefore we added 0.44 mg N L^{-1} to all model simulations of TN concentration assuming that groundwater concentrations had equilibrated with the applied wastewater nitrogen".

10. Page 4323, line 1, (1) again there is very little information on your model set up, time steps, methods used to calculate surface runoff, routing, etc… (2) here you stated that you had hourly measurements. Why compare it to daily mean simulations then? (3) Did you run your model sub-daily? Would that be an option?

- (i) In response to the lack of information on model setup, a description has been added in the Model configuration section as follows: "The DEM was used to delineate boundaries of the whole catchment and individual sub–catchments, with a stream map used to 'burn–in' channel locations to create accurate flow routings. Hourly rainfall estimates were used as hydrologic forcing data. The Penman–Monteith method (Monteith, 1965) was used to calculate evapotranspiration (ET) and potential ET. The Green and Ampt (1911) method was used to calculate infiltration, rather than the SCS curve number method. Therefore, the hourly rainfall/Green & Ampt infiltration/daily routing method (Neitsch et al., 2011) was chosen to simulate upland and in–stream processes".

(ii) The reason why those data were compared with daily mean simulations has been added in the text as follows: "The use of the high–frequency observations for model validation allowed to examine how the model performed during short (1–3 day) high flow periods".

(iii) Please see the response to *Comment #2*. We did not run sub–daily simulations, because measurements for important meteorological forcing variables (e.g., temperature, relative humidity and solar radiation) were available only at daily resolution.

11. The efficiency criteria presented in table 4 are widely known. You don't need this table. Besides, you presented what is considered as satisfactory and unsatisfactory in table 5.

- We wish to keep the efficiency criteria presented in Table 4 because the values are referred to extensively throughout the manuscript to evaluate the model fit.

e) Hydrograph and contaminant load separation

12. All the three water quality constituents are load separated with base and peak flow. Is the characteristics of all three elements the same? Are they all following the river discharge regime? Can you elaborate on that?

- The text has been added as follows: "The characteristics of concentration–discharge relationships for SS and TP are different to that for TN (Abell et al., 2013). In quick flow, there is a positive relationship between Q and concentrations of SS and TP, reflecting mobilisation of sediments and associated particulate P. Total nitrogen concentrations declined slightly in quick flow, reflecting the dilution of nitrate from groundwater".

Sensitivity analysis

13. Why log 10, why the threshold of 0.1? Page 4324, line 15, and figure 2 need a proper reference.

- Please see the response to *comment #5* regarding how the natural logarithm is now used and clarification for the threshold value of 0.2 that is chosen to decide which parameters are most sensitive.

  References have been added in Figure 2 using footnotes. Specifically: "Web–based Hydrograph Analysis Tool (Lim et al. 2005)"; Define concentrations in base flow (C_b) and quick flow (C_q) components (cf. Rimmer and Hartmann, 2014); and the natural logarithm (Krause et al., 2005)".

Results

f) Model performance

14. Please add efficiency criteria to all 8 figures in figure 3 both for calibration and validation periods. It is much easier to have them on the graphs rather than in table 5.

- Efficiency criteria are already presented in Table 5. The purpose of the graphs is to provide a visual example of model goodness–of–fit".

15. Page 4326, line 1-10.(1) Figure 4 and the explanation are very unclear. (2) The symbols used in the figure are not distinguishable. (3) I am not sure what the main point of this

paragraph and the figure is! What is the main idea of "discharge weighted daily mean concentration" and then comparing them to simulated mean? (4) What did this analysis reveal?

- (i) The caption of Figure 4 has been revised to read: "Example of a storm event showing derivation of discharge (Q)–weighted daily mean concentrations (dashed horizontal line) based on hourly measured concentrations (black dots) of suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) over two days (a–c). Comparisons of Q–weighted daily mean concentrations with simulated daily mean estimates of SS, TP and TN (scatter plot, d–f). The horizontal bars show the ranges in hourly measurements during each storm event in 2010–2012".

  (ii) In Figure 4 a–c, we removed the black horizontal line showing the simulated daily mean. No more changes were made as the symbols in the publisher's version appear to be clear.

  (iii) Please see the response to *Comment #10 (ii)*. The main point/idea was to compare discharge weighted daily mean concentration with simulated daily mean, to examine how the model performed during short (1–3 day) high flow periods.

  (iv) This analysis reveals that model uncertainty could be considerably underestimated if monthly instantaneous samples undertaken during base flow were predominantly used for model calibration. Accordingly, further text has been added to the Discussion as follows: "Most water quality data used for model calibration comprised monthly instantaneous samples taken during base flow conditions. The use of those measurements for model calibration would likely lead to considerable underestimation of constituent concentrations (notably SS and TP) due to failure to account for short–term high flow events".

16. In general the model provides poor results in water quality representation. It is always easy to blame the model not to represent the process adequately! There might be processes that are going on in the catchment and you are not including them in the model. e.g. fertilizer application… Maybe you need to revisit your conceptual model. That's exactly why you need uncertainty analysis!

- Fertilizer application was included in the model, though it is one of the inputs that will have a moderate to high level of uncertainty. We accept that there may have been

activities or processes which were not included in the input data to the model. In general we consider that we have captured the major inputs, and have added suitable text in the Discussion; please see the response to *comment #8.*

Parameter sensitivity

17. Figure 5 "Simulations for base flow and quick flow" is impossible to read. (1) You need to change the symbols. (2) There is absolutely no explanation on this figure in the text. (3) What are you trying to convey by presenting this figure? Again, what are the key points?

- (i) The symbols have been made clear and this should help to convey our main point about differentiating water quality constituents based on hydrograph separation.

  (ii) The following text has been added to the Section Results to support Fig. 5: "Model performance statistics differed between the two flow regimes (Table 6). Simulations of discharge and constituent loads under quick flow were more closely related to the measurements (i.e., higher values of R^2 and NSE) than simulations under base flow. Base flow TN load simulations during the validation period showed better model performance than simulations under quick flow. Additionally, measurements under quick flow were better reproduced by the model than the measurements for the whole simulation period. Simulations of contaminant loads matched measurements much better than for contaminant concentrations, as indicated by statistical values for model performance given in Table 5 and 6".

  (iii) Accordingly, further text has been added to the Discussion as follows: "The analysis of model performance based on datasets separated into base flow and quick flow constituents enabled uncertainties in the structure of hydrological models to be identified, denoted by different model performance between these two flow constituents".

18. Figure 7 (previous Figure 6) "Parameter sensitivity" and the corresponding text: you need to explain the method better. It is very unclear right now.

- We have revised the text for the caption of Figure 7 to read: "The standard deviation (STD) of the ln–transformed Nash–Sutcliffe efficiency (NSE) used to indicate parameter sensitivity based on one–at a–time (OAT) sensitivity analysis for separate base and quick flow components: (a) Q (discharge); (b) SS (suspended sediment); (c) MINP (mineral phosphorus); (d) NO_3–N (nitrate–nitrogen); (e) ORGN (organic nitrogen); (f) NH_4–N

(ammonium–nitrogen). A median value (0.2) derived from the STD of ln–transformed NSE was chosen as a threshold above which parameters were deemed to be 'sensitive'. Definitions of each parameter are shown in Table 3".

Discussion

g) Temporal dynamics of model performance

19. In general, I would suggest that you combine result and discussion. This way you have more space to provide more in depth analysis and you avoid repeating yourself.

- We consider that separation of the Results and Discussion provides a better and more conventional way of presenting information.

20. Page 4329, line 5, please clarify how your results show that "Our results also highlight a discrepancy between the static nature of the groundwater nitrogen pool represented in SWAT and the reality that groundwater nutrient concentrations change dynamically in a lagged response (Bain et al., 2012) to changes to sources in modified catchments".

- We have added the following on Page 4325, lines 6–9: "Modelled and measured TN concentrations were generally better aligned during base flow (Fig. 3d), apart from a mismatch prior to 1996 when monthly measured TN concentrations were substantially lower than model predictions, although the concentrations gradually increased (Fig. 3h) during the validation period (1994–1997)"; and on Page 4328, lines 23–27: "Overestimation of TN concentrations prior to 1996 reflects higher NO_3–N concentrations in groundwater during the calibration period (2004–2008) due to the wastewater irrigation operation. Nitrate concentrations appeared to reach a new quasi–steady state as wastewater loads and in–stream attenuation came into balance".

  Additional text has been added as follows: "SWAT may not adequately represent the dynamics of groundwater nutrient concentrations (Bain et al., 2012) particularly in the presence of changes in catchment inputs (e.g., with start–up of wastewater irrigation). The groundwater delay parameter was set to five years (cf. Rotorua District Council, 2006), but this did not appear to capture adequately the lag in response to increases in stream nitrate concentrations following wastewater irrigation from 1991".

21. Page 4329, line 21, is process under-representation the only reason? What about input uncertainties (for example)? That's exactly where uncertainty analysis come to play!

- We agree and have included text to indicate that uncertainty that could be contributed from uncertainties in input data or process representation in the model. Please see the response to *comment #8* for the additional text.

h) Temporal dynamics of parameter sensitivity

22. Page 4331, line 3, if you are not using SCS curve number method, why the parameter is in your calibrating parameter list then? Of course the model will be insensitive to it!

- The parameter curve number (CN2) has been removed from the parameter list in Table 3, because the Green and Ampt (1911) method was used to calculate infiltration, rather than the SCS curve number method.

23. Page 4331, line 3, "was not found to be sensitive" ! was found to be insensitive

- The corrected text reads: "The curve number (CN2) parameter was found to be insensitive in both this study and Shen et al. (2012) …".

24. It would be very interesting to see how the model performance changes in high flow and low flow while feed in different parameter set at the two stages. The main question will be then: does a temporal dynamic parameterization improve model performance? So far, you showed that the model is sensitive to different parameters in high and low flow which is also valuable.

- Yes, we did not attempt to vary the parameters with discharge. This would be a new undertaking for which in our case there may be limited data to attempt validation.

25. The title can be shortened and become more informative of the main research question.

- Please see the response *to Reviewer #2, comment #1*. The title has been revised to read: "Effects of hydrologic conditions on SWAT model performance and parameter sensitivity for a small, mixed land use catchment in New Zealand".

References:

Abbaspour, K.C.: Swat-Cup4: SWAT Calibration and Uncertainty Programs Manual Version 4, Department of Systems Analysis, Integrated Assessment and Modelling (SIAM), Eawag, Swiss Federal Institute of Aquatic Science and Technology, Duebendorf, Switzerland, pp 106, 2014.

Neyman, J.: Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability, Phil. Trans. R. Soc. A, 236, 333–380, doi:10.1098/rsta.1937.0005, 1937.

Seymour, G.: Predictive Inference: An Introduction, Chapman & Hall, New York, pp 280, 1993.

Rice, J.A.: Mathematical statistics and data analysis, Boston, MA: Cengage Learning, 2006.