Anonymous Referee #2

We thank the reviewer for the positive feedback made at the start of the general comments section. Other general comments were as followed:

“Making all the text more fluent and easy to follow, consider re-organizing a few topics of the paper…”

- ➢ In response, we have edited and re–organised a few topics of the manuscript, which are: (a) Sections 2.1 ‘Study area’ and 2.2 ‘Model configuration’ have been separated, (b) the Section ‘Parameter calibration’ has been renamed to ‘Model calibration and validation’, (c) the Section ‘Sensitivity analysis’ has been incorporated into the Section ‘Hydrograph and contaminant load separation’, (d) the Section ‘Model evaluation’ has been moved down to the end of Section 2 ‘Methods’, (e) model uncertainty analysis has been added into the Section ‘Model evaluation’, (f) a general summary has been placed at the beginning of Section 4 ‘Discussion’, (g) a new Section ‘Key uncertainties’ has been added between the two Sections ‘Temporal dynamics of model performance’ and ‘Temporal dynamics of parameter sensitivity’.

  Additional text has been added in both Sections ‘Results’ and ‘Discussion’ including 1) calibrated parameter values (have been added to Results); 2) values of model performances statistics have been added to the Results for simulations of discharge and contaminant loads, separated for the two flow regimes. Brief text has been added to the Discussion in relation to these results; 3) details of model uncertainties, based on 95% confidence intervals and 95% prediction intervals have been added to the Results; and 4) relative sensitivity analysis of parameters by randomly generating combinations of values for model parameters for each individual variable before the one–at a–time analysis of parameter sensitivities have been quantified in the Results for the separated flow constituents.

“…and also the authors should address better “the need of a robust calibration and validation, and that a calibration of a particular situation may lead to a greater uncertainty on scenario analyses”, and in this sense, it is important to clarify better how the particular case study calibration was conducted and what parameter values were obtained.”

- ➢ We have edited the Section ‘Model calibration and validation’ to provide additional details of the calibration and validation processes. We have also added the calibrated parameter values to Table 3.

"As well as, if not quantify uncertainties for this paper, but to introduce some discussion regarding the uncertainties and limitations of the methodology used, the monitored data, and separation of the hydrograph contributions (base and quick flows), and concentrations. And also pass the key findings to the reader in the end."

- To address this comment, we have added a new section to the Discussion entitled 'Key uncertainties'. This reads: "Lindenschmidt et al. (2007) found sources of uncertainty in a river water quality modelling system in terms of estimated parameter values, model input data, and model equations used to calculate processes. Model uncertainty in this study may, therefore, 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. 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 and therefore provided an uncertainty for the real values that may widely vary in representing different characteristics of the study catchment. In terms of forcing data, it appeared reasonable to assume the spring discharge rate be invariant. However, the assumption of constant values of nutrient concentrations that inadequately reflected temporal variances might be one factor causing to model uncertainty, although as a relatively minor source of model error. Most measured water quality data used for model calibration were monthly instantaneous samples taken during base flow. The use of those measurements for model calibration would lead to a considerable underestimate of constituent concentrations if the study area endures quite a high frequency of rainfall events. Inadequate representation of groundwater processes in the model structure is another key factor causing to the underestimates of model uncertainty by affecting nitrogen simulations". 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".

As described in the response to comment #20, key findings have been added in the Section 'Temporal dynamics of parameter sensitivity' in the Discussion as follows: "This study has important implications for modelling studies of similar catchments that exhibit short–term temporal fluctuations in stream flow. In particular these include small catchments with relatively steep terrain, low order streams and moderate to high rainfall".

Specific Comments:

1. The title could express better the main question and discussion of the paper;

- 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".

2. Abstract is clear and it catches the reader attention for the paper, but should also incorporate the main findings of the application on the watershed studied and possible implications;

- We have included additional text to capture the main findings of the study. Please see our response to *Referee #1, comment #2*: "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".

3. The methods section: Although the authors discuss more about the watershed's conditions on the discussion section, it would be valuable for the reader to be able to understand it before, to follow better the discussion. As what are the main processes, average precipitation,

slope, characteristics, land uses, soil types, etc. What would be typical base flow, quick flow, lateral flow contributions.

- ➢ We now provide a more detailed description of watershed characteristics in Section 2.1 ‘Study area’. Additional text is as follows: “The catchment is situated in the central North Island of New Zealand, which has a warm temperate climate. Annual mean temperature at Rotorua Airport (Fig. 1a) is 15±4 °C and annual mean evapotranspiration is 714 mm yr^{-1} (1993–2012; National Climatic Data Centre; available at http://cliflo.niwa.co.nz/). Annual mean precipitation at Kaituna rain gauge (Fig. 1a) is 1500 mm yr^{-1} (1993–2012; Bay of Plenty Regional Council). The catchment is relatively steep (mean slope = 9%; Bay of Plenty Regional Council) with predominantly pumice soils that have high macroporosity, resulting in high infiltration rates and substantial sub–surface lateral flow contributions to stream channels. Two cold–water springs (Waipa Spring and Hemo Spring) and one geothermal spring (Fig. 1b) are located in the LTS. Two cold–water springs have annual mean discharge of ~0.19 m^3 s^{-1} (Rotorua District Council) and one geothermal spring has annual mean discharge of ~0.12 m^3 s^{-1} (White et al., 2004)”.

  We note that we have already provided details of the land use composition of the catchment on Page 4320, lines 4–15, hence, no further information about land use characteristics have been included.

  After we introduced the FRI gauge on Page 4320, lines 16–21, a detailed text is added as follows: “Annual mean discharge at this site is 2.0 m^3 s^{-1} (1994–1997 and 2004–2008; Bay of Plenty Regional Council). The Puarenga Stream receives a high proportion of flow from groundwater stores and has only moderate seasonality in discharge. On average, the lowest mean daily discharge is during summer (December to February; 1.7 m^3 s^{-1}) and the highest mean daily discharge is during winter (June to August; 2.4 m^3 s^{-1})”.

4. The same goes for the SWAT model application, it is not clear for the reader, if the authors used the default configuration with default equations, or if different methods within SWAT were used. As for example, which method was used to calculate PET? Which for curve number? Which for routing? Also it is not clear in this section if the authors used the hourly input and ran SWAT with hourly data, using Green & Ampt, or if the data was aggregated on daily beforehand, and SCS method was used. Or for example what was the warm up period used? It would be important to write the chosen methods of the model in the methods section.

- ➢ In response to the reviewer’s comments, the following text has been added in the Model configuration section: “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". And in the Model calibration and validation section it has been added as follows: "One year (1993) was used for model warmup…".

5. The paper has a great amount of information for this section, as for example plant parameters, wastewater applications, etc. Tables 1 and 2 were good to concise a lot of this information. And of course this is not the main point of the paper, but it has to be sufficient for reproduction. So we advise a better description of model configuration, and also of the calibration process;

➢ Please see the section Model configuration which is now more comprehensive. Additional text has been added to this 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. Ten sub–catchments were represented in the Puarenga Stream catchment, each comprising numerous Hydrologic Response Units (HRUs). Each HRU aggregates cells with the same combination of land cover, soil, and slope. A total of 404 HRUs was defined in the model. Runoff and nutrient transport were predicted separately within SWAT for each HRU, with predictions summed to obtain the total for each sub–catchment".

6. In the calibration: (1) please cite more literature, and although the algorithm and software (SUFI-2 and SWAT-CUP) are mentioned, there is a need to explain how the calibration process was. (2) Was flow calibrated first? And then suspended sediment? And then water quality related parameters? Was it all at once? (3) Why the authors calibrated TP manually and the others with SUFI-2? (4) No Sensitivity analysis was done prior to calibration, why? What was the Objective function used?

- (i) A further reference, i.e. Wu and Chen (2015), has been added to the background text as follows: “The SUFI–2 procedure has been integrated into the SWAT Calibration and Uncertainty Program (SWAT–CUP). SUFI–2 is a procedure that efficiently quantifies and constrains parameter uncertainties/ranges from default ranges with the fewest number of iterations (Abbaspour et al., 2004), and has been shown to provide optimal results relative to the use of alternative algorithms (Wu and Chen, 2015)”.

  (ii) Parameters were calibrated in the following order: discharge (Q), SS, TP and TN. The sequence of calibration is described (Page 4322, lines 13–16) 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”.

  (iii) The reason why TP was calibrated manually is explained in the text on Page 4328, lines 14–22 as follows: “The ORGP fraction that is simulated in SWAT includes both organic and inorganic forms of particulate phosphorus, however, the representation of particulate phosphorus cycling only focusses on organic phosphorus cycling, with limited consideration of interactions between inorganic streambed sediments and dissolved reactive phosphorus in the overlying water (White et al., 2014). This contrasts with phosphorus cycling in the study stream where it has been shown that dynamic sorption processes between the dissolved and particulate inorganic phosphorus pools exert major control on phosphorus cycling (Abell and Hamilton, 2013)”.

  (iv) Sensitivity analysis was done prior to calibration using the SUFI–2 procedure. It helped to gain insight into the variances in parameter sensitivities for different flow regime components using ‘one–at a–time’ (OAT) routine. A detailed description has been added after the background of Latin hypercube sampling (LHS) 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					

7. I believe the section 2.1, 2.2 and 2.3 can be better organized. In the end of section 2.2 there is some description of the model used - SWAT, and in model evaluation a small description of calibration and validation, please revise.

➢ We have re-organised these three sections and the new structure is: 2.1 Study area, 2.2 Model configuration, 2.3 Model calibration and validation. The description of the SWAT model has been moved into Section 2.2. The description of calibration and validation has been moved into Section 2.3. The section relating to model evaluation has been moved down to the end of Section 2.

8. Table 1: (1) Please state clearly that the 15 min data was aggregated; the acronyms SS, TP and TN are in Table 1, but they were not presented in the text that is before Table1; (2) please also explain in here why there are the two validation periods with a short sentence as a footnote, for example, just to be clear. (3) Consider separating into two sections the point sources, in the contributions: spring, etc; and the abstractions, with related sources. (4) Also, why were the spring discharges constant, if there was measured data, if it was not enough for a daily series, how were they "based" on the measured data?

➢ (i) The relevant text has been adjusted in Row #2 Column #3 as follows: "FRI: 15–min stream discharge data were aggregated as daily mean values …, monthly grab samples for determination of suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) concentrations …".

(ii) A footnote has been added to Table 1 as follows: "Model validation was undertaken using two different datasets. The monthly measurements (1994–1997) were predominantly collected when base flow was the dominant contributor to stream discharge. Data from high–frequency sampling during rain events (2010–2012) were also used to validate model performance during periods when quick flow was high".

(iii) The section of point sources has been separated into two sections "Spring discharge and nutrient loads" and "Water abstraction volumes" with their relative sources.

(iv) Regarding the constant spring discharge, the flow data and nutrient concentrations were reported as mean values in the relevant sources (see Table 1). Therefore constant daily mean discharge and nutrient concentrations were assigned in this study.

9. Table1: (1) Soil characteristics, make it clear if all the SWAT needed parameters were directly from data, or how they "were determined using key physical properties" were pedo-transfer functions used? (2) Meteorological data section: include the airport station as source; (3) for the "Agricultural management practices" would be nice to subdivide to attribute what is source of what, if feasible.

➢ (i) Thanks for pointing out this inaccurate sentence. Characteristics of functional horizons from top to bottom of the soil profile (e.g. the thickness and the soil texture contents of each horizon) were derived from digital soil maps; however, the soil maps provided limited information on physical properties, a few of which were only represented as a mean value for the whole soil profile. Some other studies measured some soil property variables (e.g. saturated hydraulic conductivity) at different predetermined functional horizons, which were then used in regression analysis to estimate values for each of the horizons. This has been clarified in Table 1 as follows: "Properties were quantified based on measurements (if available) or estimated using regression analysis to estimate properties for unmeasured functional horizons".

(ii) The source of the airport station has been included in Table 1 as follows: "Rotorua Airport Automatic Weather Station, National Climate Database (available at http://cliflo.niwa.co.nz/)".

(iii) The section of Agricultural management practices has been subdivided to three sub–sections according to their relevant citations: 1) stock density (Statistics New Zealand, 2006; Ledgard and Thorrold, 1998); 2) applications of urea and di–ammonium phosphate (Statistics New Zealand, 2006; Fert Research, 2009); and 3) applications of manure–associated nutrients (Dairying Research Corporation, 1999).

10. The phrase starting with "A validation period was chosen that pre-dated the calibration period because wastewater irrigation has occurred daily since 2002, compared with weekly during the validation period (1994–1997)" in the 2.3 section is not clear, specially the "compared with weekly", please revise.

➢ We agree that this is a somewhat unusual situation that reflects issues of data availability (discharge records) and the history of management operations that are specific to this catchment.

Please see our response to Reviewer #1, comment #7. We have revised the text more clearly as follows "A validation period that pre–dated the calibration period was chosen because discharge records were available for two separate periods (1994–1997 and post 2004). In addition, the operational regime for the wastewater irrigation has varied since operations began in 1991, with a marked change occurring in 2002 when operations switched from applying the wastewater load to two blocks (rotated daily for a total of 14 blocks in a week; i.e., each block irrigated weekly), to 10–14 blocks each irrigated daily. This operational regime continues today and we therefore decided to assign the most recent (post 2002) period (2004–2008) to calibration to ensure that the model was configured to reflect current operations".

11. Calibration – (1) Table 3: Please include calibrated values, (2) and how the parameters were changed within the given range; For example was CANMX changed for all crops? (3) CN2 and slope parameters etc were changed as relative parameters, or were they changed arbitrarily within the given range? (4) Were the physical characteristics of the catchment considered, how?

➢ (i) This 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".

(ii) The parameter CANMX was not changed for all crops because the main land use in the catchment is plantation forest, therefore the value for parameter CANMX was assigned as constant for the land use type (*Pinus radiata*).

(iii) Parameters were changed by absolute values within the given ranges. The statement has been added in the text as follows: "Auto–calibrated parameters for simulations of Q, SS, and TN were changed by absolute values within the given ranges. Some of those given ranges were restricted based on the optimum values calibrated in similar studies".

Optimal parameter set was also constrained by the analysis of model uncertainty with consideration of two criteria, i.e., optimal parameter set was derived from when > 90% of measured data was bracketed by simulated output (termed P–factor) and the average thickness of the 95PPU band divided by the standard deviation of measured data (termed R–factor) was close to one. Therefore, it could avoid the homogeneity of the same model performance statistic (e.g. NSE) estimated from different parameter values that were changed by absolute values from different parameter ranges.

Regarding the manual calibration for TP simulations, we considered the information on the auto–calibrated parameter values for MINP simulations. The statement has been added in the text as follows: "Parameter values for TP simulations were manually–calibrated based on the relative percent deviation from the predetermined values of those auto–calibrated parameters for MINP simulations, given by the objective functions (e.g. NSE)".

(iv) This has been added in the text as follows: "Parameters related to the physical characteristics of the catchment were not changed because their values were considered to be representative of the catchment characteristics".

12. Calibration – Table 3: There are some parameter values here that seem very high. As for example CANMX, LAT_TTIME (1800?) etc, please revise, and justify;

➢ We have checked the values presented in this table and confirm that the values given are indeed the SWAT default ranges (Neitsch et al., 2011), as described in column heading.

13. Do we need any of these 3 formulas? Formula 1 is a weighted average; formula 2 and 3 are the same, just changing the left side, and are mass balance. Consider leaving only citation, especially since they are also on Figure 2.

➢ Equation #1 (named formula 1 by the reviewer) is necessary to keep in because it was used to calculate discharge–weighted mean concentrations based on the high–frequency measured data.

  We believe that the initial numbered Equation #5 (named formula #2 by the reviewer) is also necessary because it is central to the concept of separately considering loads associated with base flow and quick flow, which is an important focus of the study. This equation is now numbered as Eq. (2) in the manuscript.

  The initial numbered Equation #6 (named formula #3 by the reviewer) has been removed because it was rearranged from Eq. (2).

14. Figure 2 is nice, but please include the citations/sources in the figure for the methods used. Also please revise the phrase on text that calls figure 2: "Methods used to quantify parameter sensitivity…", since figure 2, explains all this methods, including the previous described separations of section 2.4;

- ➢ 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)”.

  The caption has been changed slightly to read “Figure 2. Flow chart of methods used to separate hydrograph and contaminant loads and to quantify parameter sensitivities for…”.

15. In the text of section 3.1 please cite the performance rating criteria used directly from Moriasi et al., 2007 (yes, I know Table 4 brings all information), but reading the text only should be clear the source.

- ➢ Performance rating criteria have been included in the text as follows “… model performance ratings (cf. Moriasi et al., 2007) of ‘very good’ and ‘good’ (Table 4).”

16. What about the statistics for the separated quick and base flows?

- ➢ A temporal evaluation for model performance of simulations for the separated quick and base flow components has been added. These results are now presented in Table 6, which is reproduced below. Accordingly, further text has been added to the Section Results and Discussion.

  The following text has been added to the Section Results as follows: “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”.

  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”.

Table 6. Model performance statistics for simulations of discharge (Q), and loads of suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN). Statistics were calculated for both overall and separated simulations. Q_{all} and L_{all} indicate the overall simulations; Q_{b} and L_{b} indicate the base flow simulations; Q_{q} and L_{q} indicate the quick flow simulations.

Model performance	**Statistics**	**Q**			**SS**			**TP**			**TN**		
		Q_{b}	Q_{q}	Q_{all}	L_{b}	L_{q}	L_{all}	L_{b}	L_{q}	L_{all}	L_{b}	L_{q}	L_{all}
Calibration (2004–2008)	R^2	0.84	0.84	0.77	0.66	0.68	0.61	0.24	0.65	0.39	0.72	0.97	0.95
	NSE	0.6	0.71	0.73	0.33	0.33	0.27	-6.2	0.09	-0.17	0.5	0.89	0.85
	±PBIAS%	7.5	8.7	7.8	7.57	-23.4	-3.6	45.4	40.1	43.6	0.8	6.6	2.7
Validation (1994–1997)	R^2	0.87	0.81	0.68	0.36	0.98	0.95	0.27	0.27	0.06	0.79	0.33	0.58
	NSE	0.56	0.62	0.62	-0.03	0.43	0.85	-1.9	0.04	-0.64	0.58	-0.07	0.33
	±PBIAS%	11.3	-1.2	8.8	34.5	-79.7	11.1	45.8	-9.3	37	-7.6	14.3	-2.5

R^2: coefficient of determination; NSE: Nash–Sutcliffe efficiency; PBIAS: percent bias

17. Please revise and make clearer the section 3.2. It does bring a nice discussion. Would also suggest changing the phrase: “Those sensitive flow parameters..: : :particularly sensitive”

- ➢ On reflection, we now believe that the sentence is unnecessary so we have removed it altogether.

18. Discuss why use log 10 Nash here, and not before or in both analyses?

- ➢ 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.

  The normalised format of NSE was used to rate model performance.

19. It is interesting and it would be expected that since the model was calibrated when wastewater was being applied that in the previous years used for validation the water quality components would be underestimated. But therefore a deeper discussion on the calibrated parameters may play an important role, since, are the parameters changed, so the physical meaning has also been decreased and therefore if no application is done, it underestimates, or is the model and algorithms, not replying well to different forcings? Therefore is it a limitation of the calibrated set of parameters only or/and method?

- ➢ Forcing data were changed throughout the simulation period but the parameters were not changed. Wastewater was applied during both the calibration and validation periods. However, as we discuss (from Page 4328, lines 27–29 to Page 4329, lines 1–2), “Our decision to deliberately select a validation period (1994–1997) during which the boundary conditions of the system (specifically anthropogenic nutrient loading) differed considerably from the calibration period allowed us to rigorously assess the capability of SWAT to accurately predict water quality under an altered management scenario (i.e. the purpose of most SWAT applications)”.

20. Section 4.2 is very valuable and dense, a final "closure" with key findings in the section 4.2 is advised; as maybe a small discussion of how regional the sensitivity analysis results are, or how they could be extrapolated to base flow and quick flow, it is difficult, but would be valuable.

- Additional text has been added in the Conclusions section as follows: "This study has important implications for modelling studies of similar catchments that exhibit short–term temporal fluctuations in stream flow. In particular these include small catchments with relatively steep terrain and lower order streams with moderate to high rainfall".

21. In the 4.2 section: would also like to see what is the average percentages of lateral flow to the flow contribution on the region both simulated and from local knowledge;

- Additional result has been added in Section 'Results' as follows: "Annual mean percentages of lateral flow recharge, shallow aquifer recharge and deep aquifer recharge to total water yield were predicted by SWAT as 30%, 10%, 58%, respectively".

  Additional text has also been added in Section 'Discussion' as follows: "The modelled estimates of deep aquifer recharge (58%) and combined lateral flow and shallow aquifer recharge (40%) were comparable with estimates derived by Rutherford et al. (2011), who used an alternative catchment model to derive respective estimates of 30% and 70% for these two fluxes".

Reference:

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.

Monteith, J.L.: Evaporation and the environment. In the state and movement of water in living organisms, 19th Symposia of the Society for Experimental Biology, Cambridge Univ. Press, London, U.K., 1965.

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

Rutherford, K., Palliser, C., Wadhwa, S.: Prediction of nitrogen loads to Lake Rotorua using the ROTAN model, Report prepared for Bay of Plenty Regional Council, New Zealand, 183. 2011.

Wu, H., Chen, B. 2015. Evaluating uncertainty estimates in distributed hydrological modeling for the Wenjing River watershed in China by GLUE, SUFI-2, and ParaSol methods. Ecological Engineering 76: 110–121.