Anonymous Referee #1

1. Don't use "we" and "they" in the manuscript, "the authors" has a suitable replacement for these. Revise whole of the text with this correction.

- The replacements have been made in several places, as suggested. We have avoided extensive changes and refer to the seminal text book of Day and Gastel (2012), who advocate use of the first person in paper writing to provide direct sentences in an active voice.

  Day, R., Gastel, B.: How to write and publish a scientific paper, 7th Edition, Cambridge University Press, Cambridge, UK, 2012.

2. Add some of the most important quantitative results to the Abstract.

- Quantitative results have been added in the Abstract 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)".

3. Page 4317, line 12: Change "spatial and temporal" to "spatiotemporal". Apply this for whole of the manuscript.

- No change made. The term "spatial and temporal" appeared in the Introduction when Boyle et al. (2000) is cited and the term is used only once through the whole paper. The use of "spatiotemporal" would not reflect the fact that Boyle et al. (2000) wished to consistently differentiate between spatial variation and temporal variation.

4. There are many useful and more new papers on auto-calibration in the different fields of hydrology which can increase reliability aspect of the methodology. Therefore, cite all of the below papers for this purpose:

1) Critical Areas of Iran for Agriculture Water Management According to the Annual Rainfall

2) Monthly Inflow Forecasting using Autoregressive Artificial Neural Network

3) Long-term runoff study using SARIMA and ARIMA models in the United States

4) Simulation of open- and closed-end border irrigation systems using SIRMOD

5) Analysis of potential evapotranspiration using 11 modified temperature-based models

6) A comprehensive study on irrigation management in Asia and Oceania

7) Future of agricultural water management in Africa

- No change made. We have considered each of the papers that the reviewer cites and we believe that they have limited relevance to our study; e.g., they do not refer to the model that we used (SWAT) and they do not consider water quality. Furthermore, although the papers relate to model applications that involve a calibration stage, the papers do not seem to focus on the topic of auto–calibration specifically. We are therefore unclear about which section of the manuscript the reviewer wishes us to cite these seven papers.

  The auto–calibration approach that we used has been provided with more detailed descriptions and one more literature, i.e. Wu and Chen (2015), has been cited 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)".

5. In this study, the authors measured the discharge every 15 minutes. In this case, why did the authors use daily scale instead hourly scale?

- 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".

6. The length of the calibration period is 5 years while, the length of the validation period is 4 years. This leads to increase of uncertainty, because two-third of the data is commonly applied for calibration period.

- No change made. We can cite many instances of studies which roughly balance the duration of the calibration and the validation periods, e.g., Santhi et al. (2001) and Cao et al. (2006; cited in the manuscript).

7. The data used for validation period (1994-1997) occurred before calibration data (2004-2008)! How do the authors justify this abnormal selection?

- 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. We therefore ensured that we specifically discussed the rationale for this in the original manuscript on Page 4322, lines 17–22 and Page 4328, lines 27–29 continued to Page 4329, lines 1–2.

  We have also revised the text on Page 4322, lines 17–22 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".

8. Why is there a gap between calibration and validation periods (1998-2003)? Is this due to lack of measuring? Why?

- The FRI stream–gauge, where the measurements of discharge and nutrient concentrations were undertaken, was closed in mid 1997, then re–opened late 2004 (Environment Bay of Plenty, 2007). This is described on Page 4320, lines 19 to 20: "Discharge records during 1998–2004 were intermittent and this precluded a detailed comparison of measured and simulated discharge during that period".

9. In Table 4, what is the criterion to this classification? For instance, why did the values of R-square more than 0.7 indicate a very good correlation?

- The rationale for this is explicitly stated in the caption for this table: "Performance rating criteria are based on Moriasi et al. (2007)… Moriasi et al. (2007) derived these criteria

based on extensive literature review and analysing the reported performance ratings for recommended model evaluation statistics".

10. Figure 3 underline poor performance of the model in peak and low points. I suggest to the authors to use a separate index for evaluation of the error of peak points as follows:

$$PVC = \frac{\sqrt[4]{\sum_{i=1}^{N_p}(X_i - Y_i)^2 \times (X_i)^2}}{\sqrt{\sum_{i=1}^{N_p}(X_i)^2}} \quad LVC = \frac{\sqrt[4]{\sum_{i=1}^{N_l}(X_i - Y_i)^2 \times (X_i)^2}}{\sqrt{\sum_{i=1}^{N_l}(X_i)^2}}$$

Where, Xi and Yi are the ith observed and estimated values, respectively; *X* and *Y* are the average of Xi and Yi, Np is number of peak parameter greater than one-third of the mean peak parameter observed, Nl is number of low parameter lower than one-third of the mean low parameter observed and n is the total numbers of data.

- A peak and low flow criterion (PLC) was introduced by Coulibaly et al. (2001) for ANN (artificial neural network) model evaluation. PLC was specified by two criteria. The peak value criterion (named PVC by Reviewer #1) originated from Ribeiro et al. (1998), while the low value criterion (named LVC by Reviewer #1) was modified from the PVC by Coulibaly et al. (2001).

  As suggested by the reviewer, the statistics PVC and LVC have been calculated and the values have been tabulated (see below). However, the sample sizes (N_p and N_l) are very low (1 to 10) for sediment and nutrient concentrations. Therefore, we decided not to use these statistics for model evaluation, at least for SS, TP and TN simulations.

		N_p	PVC	N_l	LVC
Q	Calibration	39	0.23	191	0.1
	Validation	53	0.29	65	0.14
SS concentration	Calibration	2	0.45	5	0.48
	Validation	1	0.56	4	0.54
TP concentration	Calibration	2	0.68	4	0.36
	Validation	1	0.80	10	0.27
TN concentration	Calibration	2	0.39	3	0.42
	Validation	2	0.24	3	0.79

11. A temporal evaluation of error indices could be useful for better understanding of performance the SWAT model. The authors can read and cite the below papers: (1) Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir; and (2) Parameters Estimate of

Autoregressive Moving Average and Autoregressive Integrated Moving Average Models and Compare Their Ability for Inflow Forecasting.

- Paper #1 and #2 (as indicated above by the reviewer) forecasted the inflow of one reservoir using two models: Auto Regression Moving Average (ARMA) and Auto Regression Integrated Moving Average (ARIMA). These two models are not used in this study; however, we agree that further consideration of how error indices vary temporally would provide valuable insight into model performance. Therefore, in addition to the combined flow statistics, we have calculated model performance statistics separately for base flow and quick flow constituents. These results are now presented in Table 6, which is reproduced below.

  The following text has been added to the 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

12. How did the authors calculate evapotranspiration as an input parameters for the SWAT model?

- This has been explained in the text as follows: "The Penman–Monteith method (Monteith, 1965) was used to calculate evapotranspiration".

13. In the Conclusion, discuss on the most important factors which are effective on variations of the base and quick flow in the study area.

- On Page 4332, lines15–17, relevant text was discussed in Conclusions as follows: "Parameters relating to main channel processes were more sensitive when estimating variables (particularly Q and SS) during base flow, while those relating to overland processes were more sensitive for simulating variables associated with quick flow".

References:

Coulibaly, P., Bobée, B., Anctil. F.: Improving extreme hydrologic events forecasting using a new criterion for ANN selection, Hydrol Process, 15, 1533–1536, doi:10.1002/hyp.445, 2001.

Monteith, J.L.: Evaporation and the environment. p. 205–234. In The state and movement of water in living organisms, XIXth Symposium. Soc. For Exp. Biol., Swansea, Cambridge University Press, 1965.

Ribeiro, J., Lauzon, N., Rousselle, J., Trung, H.T., Salas, J.D.: Comparaison de deux mode`les pour la pre´vision journalie`re en temps re´el des apports naturels, Can. J. Civil Engng 25, 291–304, 1998.

Santhi, C., Arnold, J.G., Williams, J. R., Dugas, W.A., Srinivasan, R., and Hauck, L.M.: Validation of the SWAT model on a large river basin with point and nonpoint sources, J. American Water Resources Assoc., 37, 1169-1188, 2001.

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.

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".

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.

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.

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".

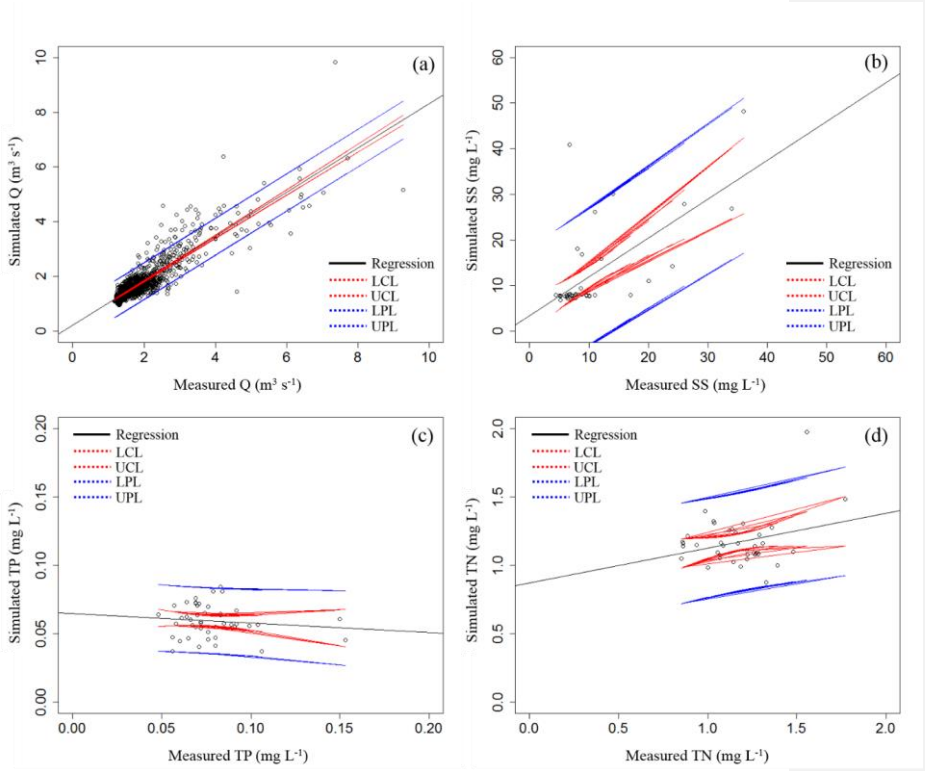


Figure 6. Regression of measured and simulated (a) discharge (Q), concentrations of (b) suspended sediment (SS), (c) total phosphorus (TP), and (d) total nitrogen (TN) including lower and upper 95% confidence limits (LCL and UCL) and lower and upper 95% prediction limits (LPL and UPL). Note that the "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.

Comments from Editor

The Article is potentially of interest for the SWAT users community and for the broader hydrological audience, but it needs significant revisions. The reviewers offer important suggestions, which I recommend to follow attentively. In addition, I provide some additional points below.

1. A plus of the paper is that it attempts to model transport in addition to flow. The results of transport simulation, however, are not so encouraging. There is a strong difference between calibration and validation performance, with much better model performance during the validation period. I would recommend a split sample approach, where the calibration and validation period are inverted, so to check the consistency of results.

- The editor stated that "model performance during the validation period showed much better than performance during the calibration period". However, the model performance ratings in Table 5 revealed that simulations of discharge and concentrations of TP and TN during model calibration indicated better model performance than validation performance, represented by statistical indices R^2, NSE, PBIAS.

  We decided not to invert the current calibration (2004–2008) and validation (1994–1997) periods into the counter way, of which the reason has been found in the response to *Reviewer #2, Comment #10*. They were also demonstrated more clearly in the text as follows: "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".

2. Can the bad performance for transport simulation during the calibration period be due to too short warmup period? What is the warmup period, and can it be increased?

- One year (1993) was used for model warmup. We believe that the length of the warmup period is not related to the poor performance of some aspects of the model. Instead, we believe that inadequate representation of groundwater processes is a key factor that affected nitrogen simulation, as we discuss in our response to *Reviewer #3, Comment #20*. 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".

3. The authors seem to compare observed instantaneous concentration data (measured once per month) with modelled monthly averages. They should compared observed averaged with simulated averages, or observed instantaneous values with simulated instantaneous values. Please clarify this aspect and correct the manuscript if necessary.

- We compared simulated daily (not monthly) mean concentrations with concentrations measured on respective days. The measured data are 'instantaneous' in that they relate to grab samples that were collected at monthly frequency. In addition, we also compare simulated daily mean concentrations with discharge–weighted mean daily concentrations that were calculated based on samples collected every 1–2 h during high flow events. These measurements are more representative of 'real' daily mean values than single instantaneous samples collected during separate days. Thus, a key focus of our paper is to examine the uncertainties that are associated with using concentration data that are infrequent relative to discharge to calibrate hydrologic models of small catchments; something that is common practice in catchment modelling. This has been clarified as we discuss in our response to *Reviewer #3, Comment #7 (i)*: "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".

4. The paper structure could be improved. (1) "study area and model configuration" should be 2 separate paragraphs. (2) The model configuration section needs more details. E.g. how many HRUs does the catchment have? How were they defined? (3) How many parameters in total?, etc.

- Thanks for the suggestion.

(i) Please see the response to Reviewer #2, Comment #7 that Sections 2.1 'Study area' and 2.2 'Model configuration' have been separated.

(ii) Please see the response to Reviewer #2, Comment #5 that the section Model configuration is now more comprehensive 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".

(iii) There were a total of 197 parameters involved for the model configuration of this study. This has added (see Model configuration) in the text: "There were a total of 197 model parameters. Values of SWAT parameters were assigned based on…".

5. Tables 2 and 3: can the parameters corresponding to hydrology, chemistry and sediment transport simulations be clearly separated.

- Table 2 shows prior–estimated parameter values for three dominant types of land–cover in the Puarenga Stream catchment.

  Table 3 can be separated for discharge and sediment in more details of their exclusive parameters. Please see a revised version below. Phosphorus and nitrogen parameters have been separated already.

Parameter	Definition	Unit	Default range
Q			
EVRCH.bsn	Reach evaporation adjustment factor		0.5–1
SURLAG.bsn	Surface runoff lag coefficient		0.05–24
ALPHA_BF.gw	Base flow alpha factor (0–1)		0.0071–0.0161
GW_DELAY.gw	Groundwater delay	d	0–500
GW_REVAP.gw	Groundwater “revap” coefficient		0.02–0.2
GW_SPYLD.gw	Special yield of the shallow aquifer	$m^3\ m^{-3}$	0–0.4
GWHT.gw	Initial groundwater height	m	0–25
GWQMN.gw	Threshold depth of water in the shallow aquifer required for return flow to occur	mm	0–5000
RCHRG_DP.gw	Deep aquifer percolation fraction		0–1
REVAPMN.gw	Threshold depth of water in the shallow aquifer required for “revap” to occur	mm	0–500
CANMX.hru	Maximum canopy storage	mm	0–100
EPCO.hru	Plant uptake compensation factor		0–1
ESCO.hru	Soil evaporation compensation factor		0–1
HRU_SLP.hru	Average slope steepness	$m\ m^{-1}$	0–0.6
LAT_TTIME.hru	Lateral flow travel time	d	0–180
RSDIN.hru	Initial residue cover	$kg\ ha^{-1}$	0–10000
SLSOIL.hru	Slope length for lateral subsurface flow	m	0–150
CH_K2.rte	Effective hydraulic conductivity in the main channel alluvium	$mm\ h^{-1}$	0–500
CH_N2.rte	Manning's N value for the main channel		0–0.3
CH_K1.sub	Effective hydraulic conductivity in the tributary channel alluvium	$mm\ h^{-1}$	0–300
CH_N1.sub	Manning's N value for the tributary channel		0.01–30
SS			
CH_COV1.rte	Channel erodibility factor		0–0.6
CH_COV2.rte	Channel cover factor		0–1
LAT_SED.hru	Sediment concentration in lateral flow and groundwater flow	$mg\ L^{-1}$	0–5000
PRF.bsn	Peak rate adjustment factor for sediment routing in the main channel		0–2

SPCON.bsn	Linear parameter for calculating the maximum amount of sediment that can be re–entrained during channel sediment routing		0.0001–0.01
SPEXP.bsn	Exponent parameter for calculating sediment re–entrained in channel sediment routing		1–1.5
OV_N.hru	Manning's N value for overland flow		0.01–30
SLSUBBSN.hru	Average slope length	m	10–150
USLE_P.mgt	USLE equation support practice factor		0–1

~~Modelling water, sediment and nutrient fluxes from a mixed land use catchment in New Zealand:~~ Effects of hydrologic conditions on SWAT model performance and parameter sensitivity for a small, mixed land use catchment in New Zealand

Commented [MW1]: Reviewer #2, Comment #1
Reviewer #3, Comment #25

W. Me[1,2], J. M. Abell[1,*], D. P. Hamilton[1]

[1]{Environmental Research Institute, University of Waikato, Private Bag 3105, Hamilton 3240, New Zealand}

[2]{College of Hydrology and Water Resources, Hohai University, Nanjing, 210098, People's Republic of China}

[*]{now at: Ecofish Research Ltd., Suite 1220 – 1175 Douglas Street, Victoria, British Columbia, Canada}

Correspondence to: W. Me (yaowang0418@gmail.com)

Abstract

The Soil Water Assessment Tool (SWAT) was configured for the Puarenga Stream catchment (77 km^2), Rotorua, New Zealand. The catchment land use is mostly plantation forest, some of which is spray–irrigated with treated wastewater. A Sequential Uncertainty Fitting (SUFI–2) procedure was used to auto–calibrate unknown parameter values in the SWAT model. ~~which was applied to the Puarenga catchment. Discharge, sediment, and nutrient variables~~loads were ~~then partitioned into two components (base flow and quick flow) based on hydrograph separation. A manual procedure (one–at a–time sensitivity analysis) was then used to quantify parameter sensitivity for the two hydrologically separated regimes.~~ Model validation was performed using two datasets: 1) monthly instantaneous measurements of suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) concentrations; and 2) high–frequency (1–2 h) data measured during rainfall events. 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 discharge, sediment SS, TP and nutrient TN concentrations with daily mean discharge–weighted high–frequency, event based 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). allowed the error in model predictions to be quantified. This comparison highlighted the potential for model error associated with quick–flow fluxes in flashy lower–order streams to be underestimated compared with low–frequency (e.g. monthly) measurements derived predominantly from base flow measurements. To address this, To overcome this problem wWe advocaterecommend that high–frequency, event–based monitoring data are used to support calibration and validation.the use of high–frequency, event–based monitoring data during calibration and dynamic parameter values with some dependence on discharge regime. Simulated discharge, SS, TP and TN loads were partitioned into two components (base flow and quick flow) based on hydrograph separation. A manual procedure (one–at a–time sensitivity analysis) was used to quantify parameter sensitivity for the two hydrologically–separated regimes. 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. 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.quantifying uncertainty in hydrological models, particularly for studies where model simulations are used to simulate responses of stream discharge and composition to changes in irrigation and land management.

Commented [MW2]: Reviewer #3, Comment #1

Commented [MW3]: Reviewer #1, Comment #2; Reviewer #2, Comment #2; Reviewer #3, Comment #4

Commented [MW4]: Reviewer #3, Comment #2 (ii)

Commented [MW5]: Reviewer #2, Comment #2

Commented [MW6]: Reviewer #3, Comment #3

1 Introduction

Catchment models are valuable tools for understanding natural processes occurring at basin scales and for simulating the effects of different management

regimes on soil and water resources (e.g. Cao et al., 2006). Model applications may have uncertainties as a result of errors associated with the forcing variables, measurements used for calibration, and conceptualisation of the model itself (Lindenschmidt et al., 2007). The ability of catchment models to simulate hydrological processes and pollutant loads can be assessed through analysis of uncertainty or errors during a calibration process that is specific to the application domain (White and Chaubey, 2005).

The Soil and Water Assessment Tool (SWAT) model is increasingly used to predict discharge, sediment and nutrient loads on a temporally resolved basis, and to quantify material fluxes from a catchment to the downstream receiving environment such as a lake (e.g. Nielsen et al., 2013). The SWAT model is physically–based and provides distributed descriptions of hydrologic processes at sub–basin scale (Arnold et al., 1998; Neitsch et al., 2011). It has numerous parameters, some of which can be fixed on the basis of pre–existing catchment data (e.g. soil maps) or knowledge gained in other studies. However, values for other parameters need to be assigned during a calibration process as a result of complex spatial and temporal variations that are not readily captured either through measurements or within the model algorithms themselves (Boyle et al., 2000). Such parameter values assigned during calibration are therefore lumped, i.e., they integrate variations in space and/or time and thus provide an approximation for real values which often vary widely within a study catchment. Model calibration is an iterative process whereby parameters are adjusted to the system of interest by refining model predictions to fit closely with observations under a given set of conditions (Moriasi et al., 2007). Manual calibration depends on the system used for model application, the experience of the modellers, and knowledge of the model algorithms. It tends to be subjective and time–consuming. By contrast, auto–calibration provides a less labour–intensive approach by using optimisation algorithms (Eckhardt and Arnold, 2001). The Sequential Uncertainty Fitting (SUFI–2) procedure has previously been applied to auto–calibrate discharge parameters in a SWAT application for the Thur River, Switzerland (Abbaspour et al., 2007), as well as for groundwater recharge, evapotranspiration and soil storage water considerations in West Africa (Schuol et al., 2008). Model validation is subsequently performed using measured data that are independent of those used for calibration (Moriasi et al., 2007).

Values for hydrological parameter values in the SWAT model can vary temporally. Cibin et al. (2010) found that the optimum calibrated values for hydrological parameters varied with different flow regimes (low, medium and high), thus suggesting that SWAT model performance can be optimised by assigning parameter values based on hydrological characteristics. Other work has similarly demonstrated benefits from assigning separate parameter values to low, medium, and high discharge periods (Yilmaz et al., 2008), or based on whether a catchment is in a dry, drying, wet or wetting state (Choi and Beven, 2007). Such temporal dependence of model parameterisation on hydrologic conditions has implications for model performance. Krause et al. (2005) compared different statistical metrics of hydrological model performance separately for base–flow periods and storm events to evaluate the performance. ~~They~~ The authors 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 simulations during storm events, while the relative form of NSE was better for base flow periods. Similarly, Guse et al. (2014) investigated temporal dynamics of sensitivity of hydrological parameters and SWAT model performance using Fourier amplitude sensitivity test (Reusser et al., 2011) and cluster analysis (Reusser et al., 2009). The authors ~~They~~ found that three groundwater parameters were highly sensitive during quick flow, while one evaporation parameter was most sensitive during base flow, and model performance was also found to vary significantly for the two flow regimes. Zhang et al. (2011) calibrated SWAT hydrological parameters for periods separated on the basis of six climatic indexes. Model performance improved when different values were assigned to parameters based on six hydroclimatic periods. Similarly, Pfannerstill et al. (2014) found that assessment of model performance was improved by considering an additional performance statistic for very low–flow simulations amongst five hydrologically–separated regimes.

To date, analysis of temporal dynamics of SWAT parameters has predominantly focussed on simulations of discharge rather than water quality constituents. This partly reflects the paucity of comprehensive water quality data for many catchments; near–continuous discharge data can readily be collected but this is not the case for water quality parameters such as suspended sediment or nutrient concentrations. Data collected in monitoring programmes that involve

Commented [MW7]: Reviewer #2, Comment #18
Reviewer #3, Comment #5

Commented [MW8]: Reviewer #1, Comment #1

sampling at regular time intervals (e.g. monthly) are often used to calibrate water quality models, but these are unlikely to fully represent the range of hydrologic conditions in a catchment (Bieroza et al., 2014). In particular, water quality data collected during storm–flow periods are rarely available for SWAT calibration, thus prohibiting opportunities to investigate how parameter sensitivity varies under conditions which can contribute disproportionately to nutrient or sediment transport, particularly in lower–order catchments (Chiwa et al., 2010; Abell et al., 2013). Failure to fully consider storm–flow processes could therefore result in overestimation of model performance. Thus, further research is required to examine how water quality parameters vary during different flow regimes and to understand how model uncertainty may vary under future climatic conditions that affect discharge regimes (Brigode et al., 2013).

In this study, the SWAT model was configured to a relatively small, mixed land use catchment in New Zealand that has been the subject of an intensive water quality sampling programme designed to target a wide range of hydrologic conditions. A catchment–wide set of parameters was calibrated using the SUFI–2 procedure which is integrated into the SWAT Calibration and Uncertainty Program (SWAT–CUP). The objectives of this study were to: (1) quantify the performance of the model in simulating discharge and fluxes of suspended sediments and nutrients at the catchment outlet; (2) rigorously evaluate model performance by comparing daily simulation output with monitoring data collected under a range of hydrologic conditions; and (3) quantify whether parameter sensitivity varies between base flow and quick flow conditions.

2 Methods

2.1 Study area ~~and model configuration~~

Commented [MW9]: Reviewer #2, Comment #7

The Puarenga Stream is the second–largest surface inflow ~~($2.03\ m^3\ s^{-1}$)~~ to Lake Rotorua (Bay of Plenty, New Zealand) and drains a catchment of 77 km^2. 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 haves 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).

The predominant land use (47%) is exotic forest (*Pinus radiata*). Approximately 26% is managed pastoral farmland, 11% mixed scrub and 9% indigenous forest. Since 1991, treated wastewater has been pumped from the Rotorua Wastewater Treatment Plant and spray–irrigated over 16 blocks of total area of 1.93 km^2 in the Whakarewarewa Forest (Fig. 1a). Following this, it took approximately four years before elevated nitrate concentrations were measured in the receiving waters of the Puarenga Stream (Lowe et al., 2007). Prior to 2002, the irrigation schedule entailed applying wastewater to two blocks per day so that each block was irrigated approximately weekly. Since 2002, 10 to 14 blocks have been irrigated simultaneously at daily frequency. Over the entire period of irrigation, nutrient concentrations in the irrigated water have gradually decreased as improvements in treatment of the wastewater have been made (Lowe et al., 2007).

Measurements from the Forest Research Institute (FRI) stream–gauge (1.7 km upstream of Lake Rotorua; Fig. 1b) were considered representative of the downstream/outlet conditions of the Puarenga Stream. The FRI stream–gauge was closed in mid 1997, then reopened late in 2004 (Environment Bay of Plenty, 2007). 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}). Discharge records during 1998–2004 were intermittent and this precluded a detailed comparison of measured and simulated discharge during that period. In July 2010, the gauge was repositioned 720 m downstream to the State Highway 30 (SH 30) bridge (Fig. 1b).

Commented [MW10]: Reviewer #2, Comment #3

Commented [MW11]: Reviewer #2, Comment #3

Commented [MW12]: Reviewer #1, Comment #8

2.2 Model configuration

SWAT input data requirements included a digital elevation model (DEM), meteorological records, records of springs and water abstraction, soil characteristics, land use classification, and management schedules for key land uses (pastoral farming, wastewater irrigation, and timber harvesting). The SWAT model version used (SWAT2009_rev488) runs on a daily time step.

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 ~~by~~within SWAT for each HRU, with predictions summed to obtain the total for each sub–catchment.

Descriptions and sources of the data used to configure the SWAT model are given in Table 1. There were a total of 197 model parameters. Values of SWAT ~~required~~ parameters were assigned based on: i) measured data (e.g. some~~most~~ 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~~ed~~ where parameters were not otherwise prescribed ~~values if other information was lacking~~.

SWAT simulates loads of 'mineral phosphorus' (MINP) and 'organic phosphorus' (ORGP) of which the sum is total phosphorus (TP). The MINP fraction represents soluble P either in mineral or in organic form, while ORGP refers to particulate P bound either by algae or by sediment (White et al., 2014). Soluble P may be ~~up~~taken up during algae growth, or ~~be~~released from benthic sediment. Either fraction can be transformed to particulate P contained in algae or sediment.

Commented [MW13]: Reviewer #2, Comment #7

Commented [MW14]: Reviewer #1, Comment #5
Reviewer #3, Comment #2 (iii)

Commented [MW15]: Reviewer #1, Comment #12

Commented [MW16]: Reviewer #2, Comment #4

Commented [MW17]: Reviewer #3, Comment #10 (i)

Commented [MW18]: Reviewer #2, Comment #5

Commented [MW19]: Editor, Comment #4 (iii)

Commented [MW20]: Reviewer #3, Comment #7 (ii)

SWAT simulates loads of nitrate–nitrogen (NO_3–N), ammonium–nitrogen (NH_4–N) and organic nitrogen (ORGN), the sum of which is total nitrogen (TN). Nitrogen parameters were auto–calibrated for each N fraction. The SWAT model does not account for the initial nitrate concentration in shallow aquifers, ~~an issue~~ as also noted by Conan et al. (2003). Ekanayake and Davie (2005) indicated that SWAT underestimated N loading from groundwater and suggested a modification by adding a background concentration of nitrate in streamflow to represent groundwater nitrate contributions. 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 ~~W~~we added 0.44 mg N L^{-1} to all model ~~estimate~~simulations of TN concentration assuming that groundwater concentrations had equilibrated with the applied wastewater nitrogen.

~~, based on groundwater composition data from Paku (2001).~~

2.3 ~~Parameter~~ Model calibration and validation

~~Unknown parameter values (Table 3) were assigned based on either automated or manual calibration. Manual calibration was undertaken for 11 parameters related to total phosphorus (TP), while a Sequential Uncertainty Fitting (SUFI-2) procedure was applied to auto-calibrate 31 parameters for simulations of discharge and suspended sediment (SS), and 17 parameters related to total nitrogen (TN). Discharge measured every 15 minutes and water quality data collected monthly by Bay of Plenty Regional Council at the FRI stream gauge (Fig. 1b), were used for model evaluation.~~ Daily mean discharge ~~(from 15-minute measurements)~~ was firstly ~~compared~~calibrated based on ~~with~~ daily mean ~~simulated discharge~~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. ~~Concentrations of SS, TP and TN measured monthly were compared with the respective simulated monthly values (derived from daily mean outputs).~~ One year (1993) was used for model warmup. The calibration period

Commented [MW21]: Reviewer #3, Comment #9 (iii)

Commented [MW22]: Reviewer #2, Comment #7

Commented [MW23]: Reviewer #2, Comment #7

Commented [MW24]: Reviewer #2, Comment #6 (ii) Reviewer #3, Comment #7 (i)

Commented [MW25]: Reviewer #2, Comment #4

was from 2004 to 2008 and the validation period was from 1994 to 1997. A validation period ~~was chosen~~ 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~~because~~.

~~wastewater irrigation has occurred daily since 2002, compared with weekly during the validation period (1994–1997). Therefore, because the groundwater nutrient pool is not dynamically modelled in SWAT, we chose to calibrate the model to reflect current operations so that it can later be used to examine how changes to land management may affect current water quality.~~

~~Unknown p~~Parameter values ~~values~~ that were not derived from measurements or the literature ~~(Table 3)~~ were assigned based on either automated or manual calibration (Table 3). Manual calibration was undertaken for 11 parameters related to ~~total phosphorus (~~TP~~)~~, while a Sequential Uncertainty Fitting (SUFI–2) procedure was applied to auto–calibrate ~~31~~21 parameters for ~~simulations of~~ discharge simulations,~~and~~ nine parameters for SS simulations~~suspended sediment (SS)~~, and 17 parameters related to ~~total nitrogen (~~TN~~)~~. 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). SUFI–2 involves Latin hypercube sampling (LHS), which is a method that ~~efficiently quantifies and constrains parameter uncertainties from default ranges with the fewest number of iterations. It~~ generates a sample of plausible parameter values from a multidimensional distribution and ~~is widely applied in uncertainty analysis~~ ensures that samples cover the entire parameter space, therefore ensuring that the optimum solution is not a local minimum (Marino et al., 2008).

Commented [MW26]: Reviewer #1, Comment #7
Reviewer #2, Comment #10

Commented [MW27]: Reviewer #2, Comment #7

Commented [MW28]: Reviewer #1, Comment #4
Reviewer #2, Comment #6 (i)

Commented [MW29]: Reviewer #3, Comment #6

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.

Commented [MW30]: Reviewer #2, Comment #6 (iv)
Reviewer #3, Comment #7 (iii)

SUFI–2 considers two criteria to constrain uncertainty in each iteration. 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. 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).

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 'optimal' parameter value was obtained when the Nash–Sutcliffe efficiency (NSE) criterion was satisfied (NSE > 0.5; Moriasi et al., 2007). 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 optimalum values calibrated in similar studies. Parameter values for TP simulations were manually–calibrated based ona the relative percent deviation from the predetermined values of those auto–calibrated parameters for MINP simulations, given by the objective functions (e.g. NSE). Parameters related to the physical characteristics of the catchment were not changed because their values were considered to be representative of the catchment characteristics.

~~Subsequent iterations were undertaken to produce narrower parameter ranges. Optimal parameter values were considered to occur when > 90% of measured data was bracketed by simulated output and the R–factor was close to one. Spatial distribution of parameters was not considered in this study as a result of the small study area size (77 km^2). Steps in the SUFI–2 application are outlined by Abbaspour et al. (2004) who integrated the SUFI–2 procedure into the SWAT Calibration and Uncertainty program (SWAT–CUP) and linked SWAT–CUP to the SWAT model.~~

~~SWAT simulates loads of 'mineral phosphorus' (MINP) and 'organic phosphorus' (ORGP) of which the sum is total phosphorus (TP). The MINP fraction represents soluble P either in mineral or in organic form, while ORGP refers to particulate P bound either by algae or by sediment (White et al., 2014). Soluble P may be uptaken during algae growth, or be released from benthic sediment. Either fraction can be transformed to particulate P contained in algae or sediment.~~

~~SWAT simulates loads of nitrate–nitrogen (NO_3–N), ammonium–nitrogen (NH_4–N) and organic nitrogen (ORGN), the sum of which is total nitrogen (TN). Nitrogen parameters were auto–calibrated for each N fraction. The SWAT model does not account for the initial nitrate concentration in shallow aquifers, an issue~~

Commented [MW31]: Reviewer #3, Comment #7 (iv)

Commented [MW32]: Reviewer #3, Comment #7 (v)

Commented [MW33]: Reviewer #3, Comment #7 (vi)

Commented [MW34]: Reviewer #2, Comment #11 (iii)

Commented [MW35]: Reviewer #2, Comment #11 (iv)

~~also noted by Conan et al. (2003). Ekanayake and Davie (2005) indicated that SWAT underestimated N loading from groundwater and suggested a modification by adding a background concentration of nitrate in streamflow to represent groundwater nitrate contributions. We added 0.44 mg N L^{-1} to all model estimates of TN concentration, based on groundwater composition data from Paku (2001).~~

~~2.3 Model evaluation~~

~~Discharge measured every 15 minutes and water quality data collected monthly by Bay of Plenty Regional Council at the FRI stream gauge (Fig. 1b), were used for model evaluation. Daily mean discharge (from 15–minute measurements) was compared with daily mean simulated discharge. Concentrations of SS, TP and TN measured monthly were compared with the respective simulated monthly values (derived from daily mean outputs). The calibration period was from 2004 to 2008 and the validation period was from 1994 to 1997. 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). Therefore, because the groundwater nutrient pool is not dynamically modelled in SWAT, we chose to calibrate the model to reflect current operations so that it can later be used to examine how changes to land management may affect current water quality.~~

In addition, high–frequency (1–2 h) water quality sampling was undertaken at the FRI stream–gauge during 2010–2012 to derive estimates of daily mean contaminant loads during storm events. Samples were analysed for SS (nine events), TP and TN (both 14 events) over sampling periods of 24–73 h. The sampling programme was designed to encompass pre–event base flow, storm generated quick flow and post–event base flow (Abell et al., 2013). These data permitted calculation of daily discharge–weighted (Q–weighted) mean concentrations to compare with modelled daily mean estimates. 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. The Q–weighted mean concentrations C_{QWM} were calculated as:

$$C_{\mathrm{QWM}} = \frac{\sum_{i=1}^{n} C_i Q_i}{\sum_{i=1}^{n} Q_i} \quad (1)$$

Commented [MW36]: Reviewer #2, Comment #7

Commented [MW37]: Reviewer #3, Comment #10 (ii) and #15 (iii)

Commented [MW38]: Reviewer #3, Comment #2 (i)

where n is number of samples, C_i is contaminant concentration measured at time i, and Q_i is discharge measured at time i.

~~Model goodness of fit was assessed graphically and quantified using coefficient of determination (R^2), Nash Sutcliffe efficiency (NSE) and percent bias (PBIAS; Table 4). R^2 (range 0 to 1) and NSE (range $-\infty$ to 1) values are commonly used to evaluate SWAT model performance at daily time step (Gassman et al., 2007). PBIAS value indicates the average tendency of simulated outputs to be larger or smaller than observations (Gupta et al., 1999).~~

Commented [MW39]: Reviewer #2, Comment #7

2.4 Hydrograph and contaminant load separation

The Web–based Hydrograph Analysis Tool (Lim et al., 2005) was applied to partition both measured and simulated discharges into base flow (Q_b) and quick flow (Q_q). An Eckhardt filter parameter of 0.98 and ratio of base flow to total discharge of 0.8 were assumed (cf. Lim et al., 2005). There were a total of 60 days without quick flow during the calibration period (2004–2008) and 1379 days for which hydrograph separation defined both base flow and quick flow.

Contaminant (SS, TP and TN) concentrations (C_{sep}) were partitioned into base flow (C_b') and quick flow components (C_q'; cf. Rimmer and Hartmann, 2014) to separately examine the sensitivity of water quality parameters during base flow and quick flow:

$$C_{sep} = \frac{Q_q \times C_q' + Q_b \times C_b'}{Q_q + Q_b} \quad (~~5~~2)$$

C_b' for each contaminant was estimated as the average concentration for the 60 days with no quick flow. C_q' for each contaminant was calculated by rearranging Eq. (~~5~~2). ~~as:~~

~~$$\overline{C_q'} = \frac{(Q_q + Q_b) \times \overline{C_{sep}} - Q_b \times \overline{C_b'}}{Q_q} \quad (6)$$~~

To ~~retain Eq. (6) rational,~~ensure that C_q' is ~~must be~~ positive, ~~therefore~~ C_b' is constrained to be the minimum ~~between~~ of $\overline{C_{sep}}$ and C_{sep}. Measured and simulated base flow and quick flow contaminant loads were then calculated.

~~2.5 Sensitivity analysis~~

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. OAT sensitivity analysis was then employed by varying the parameter of interest among ten equidistant values within the default range. 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 standard deviation (STD) of log_{10} transformed NSE values was calculated from the sensitivity analysis for each variable and for the two flow regimes (base flow and quick flow).~~

Parameters were ranked from most to least sensitive on the basis of the sensitivity metric (STD of ~~log_{10}~~ln–transformed NSE), using a value of 0.~~1~~2 as a threshold above which parameters were deemed particularly 'sensitive'. 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. Methods used to separate the two flow constituents and to quantify parameter sensitivity are illustrated in Fig. 2.

~~2.3~~5 Model evaluation

Model goodness–of–fit was assessed graphically and quantified using coefficient of determination (R^2), Nash–Sutcliffe efficiency (NSE) and percent bias (PBIAS; Table 4). R^2 (range 0 to 1) and NSE (range -∞ to 1) values are commonly used to evaluate SWAT model performance at daily time step (Gassman et al., 2007). PBIAS value indicates the average tendency of simulated outputs to be larger or smaller than observations (Gupta et al., 1999).

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 R software was used to graphically show the 95% confidence and prediction intervals for measurement data (Neyman, 1937) and

Commented [MW40]: Reviewer #2, Comment #13

Commented [MW41]: Reviewer #2, Comment #6 (iv)
Reviewer #3, Comment #7 (iii)

Commented [MW42]: Reviewer #2, Comment #18
Reviewer #3, Comment #5 and #13

Commented [MW43]: Reviewer #3, Comment #5 and #13

Commented [MW44]: Reviewer #2, Comment #7

model prediction intervals (Seymour, 1993) for Q and concentrations of SS, TP and TN during the calibration period (2004–2008).

Commented [MW45]: Reviewer #3, Comment #8, #16, and #21

3 Results

3.1 Model performance and uncertainty

Commented [MW46]: Reviewer #3, Comment #8, #16, and #21

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).

Commented [MW47]: Reviewer #3, Comment #7 (v)

The parameters that provided the best statistical outcomes (i.e, best match to observed data) are given in Table 3. 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). Model uncertainties for simulations of Q and SS, TP and TN concentrations are shown in Fig. 6.

Commented [MW48]: Reviewer #2, Comment #11 (i)
Reviewer #3, Comment #7 (vii)

Commented [MW49]: Reviewer #3, Comment #8, #16, and #21

Modelled and measured base flow showed high correspondence, although measured daily mean discharge during storm peaks was often underestimated (Fig. 3a and 3e). 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. Modelled SS concentrations overestimated measurements of monthly grab samples by an average of 18.3% during calibration and 0.32% during validation (Fig. 3b and 3f). Measured TP concentrations in monthly grab samples were underestimated by 23.8% during calibration (Fig. 3c) and 24.5% during validation (Fig. 3g). Similarly, measured TP loads were underestimated by 34.5% and 38.4%, during calibration and validation, respectively. 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 ~~they~~ the concentrations gradually increased (Fig. 3h) during the validation period (1994–1997). The average measured TN load increased from 134 kg N d^{-1} prior to 1996, to 190 kg N d^{-1} post 1996. The comparable increase in modelled TN load was 167 kg N d^{-1} to 205 kg N d^{-1}, respectively.

Commented [MW50]: Reviewer #2, Comment #21

Statistical evaluations of goodness–of–fit are shown in Table 5. The R^2 values for discharge were 0.77 for calibration and 0.68 for validation, corresponding to model performance ratings (cf. Moriasi et al., 2007) of 'very good' and 'good' (~~cf.~~ Table 4). Similarly, the NSE values for discharge were 0.73 (good) for calibration and 0.62 (satisfactory) for validation. Positive PBIAS (7.8% for calibration and 8.8% for validation) indicated a tendency for underestimation of daily mean discharge, however, the low magnitude of PBIAS values corresponded to a performance rating of 'very good'. The R^2 values for SS were 0.42 (unsatisfactory) for calibration and 0.80 for validation (very good). Similarly, the NSE values for SS were -0.08 (unsatisfactory) for calibration and 0.76 (very good) for validation. The model did not simulate trends well for monthly measured TP and TN concentrations. The R^2 values for TP and TN were both < 0.1 (unsatisfactory) during calibration and validation and NSE values were both < 0 (unsatisfactory). Values of PBIAS corresponded to 'good' or 'very good' performance ratings for TP and TN.

Observed Q–weighted daily mean concentrations derived from hourly measurements and simulated daily mean concentrations of SS, TP and TN during an example two–day storm event are shown in Fig. 4a–4c. The simulation of SS and TN concentrations was somewhat better than for TP. Comparisons of Q–weighted daily mean concentrations (C_{QWM}) during storm events from 2010 to 2012 are shown in Fig. 4d–4f for SS (nine events), TP and TN (both 14 events). The C_{QWM} of TP exceeded the simulated daily mean by between 0.02 and 0.2 mg P L^{-1}, and on average, the model underestimated measurements by 69.4% (Fig. 4e). Although R^2 and NSE values for C_{QWM} of TN were unsatisfactory (Table 5), they were both close to the threshold for satisfactory performance (0.5). For C_{QWM} of SS and TP, R^2 and NSE values indicated that the model performance was unsatisfactory. The PBIAS value of -0.87 for C_{QWM} of TN corresponded to model performance ratings of 'very good', while the PBIAS values for C_{QWM} of SS and TP were 43.9 and 69.4, respectively, indicating satisfactory model performance.

Measured and simulated discharge and contaminant ~~concentration~~loads separated for the two flow regimes (base flow and quick flow)~~,~~ are shown in Fig. 5. Model performance statistics differed between the two flow regimes (Table 6). Simulations of discharge and constituent loads under quick flow were more

Commented [MW51]: Reviewer #2, Comment #15

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.

Commented [MW52]: Reviewer #1, Comment #11
Reviewer #2, Comment #16
Reviewer #3, Comment #17 (ii)

3.2 Separated ~~P~~parameter sensitivity

~~Measured and simulated discharge and contaminant concentrations for the two flow regimes (base flow and quick flow), are shown in Fig. 5.~~

Based on the ranking of relative sensitivities of hydrological and water quality parameters derived from the SUFI–2 procedure (see Table 7), ~~T~~the OAT sensitivity analysis undertaken separately for base flow and quick flow identified three parameters that most influenced the quick flow estimates, and five parameters that most influenced the base flow estimates (parameters above the dashed line in Fig. ~~6a~~7a). ~~Those sensitive flow parameters specifically relate to the relevant flow components, providing a mechanistic basis for the finding that they were particularly sensitive.~~ Channel hydraulic conductivity (CH_K2) is used to estimate the peak runoff rate (Lane, 1983). Lateral flow slope length (SLSOIL) and lateral flow travel time (LAT_TIME) have an important controlling effect on the amount of lateral flow entering the stream reach during quick flow. Both slope (HRU_SLP) and soil available water content (SOL_AWC) were particularly sensitive for the base flow simulation because they affect lateral flow within the kinematic storage model in SWAT (Sloan and Moore, 1984). The aquifer percolation coefficient (RCHRG_DP) and the base flow alpha factor (ALPHA_BF) strongly influenced base flow calculations (Sangrey et al., 1984), as did the channel Manning's N value (CH_N2) which is used to estimate channel flow (Chow, 2008).

Commented [MW53]: Reviewer #2, Comment #6 (iv)
Reviewer #3, Comment #7 (iii)

Commented [MW54]: Reviewer #2, Comment #17

For SS loads, 12 and four parameters, respectively, were identified as sensitive in relation to the simulations of base flow and quick flow (parameters above the dashed line in Fig. ~~6b~~7b). Parameters that control main channel processes (e.g. CH_K2 and CH_N2) and subsurface water transport processes (e.g.

LAT_TIME and SLSOIL) were found to be much more sensitive for base flow SS load estimations. Exclusive parameters for SS estimations, such as SPCON (linear parameter), PRF (peak rate adjustment factor), SPEXP (exponent parameter), CH_COV1 (channel erodibility factor), and CH_COV2 (channel cover factor) were found to be much more sensitive in base flow SS load, while LAT_SED (SS concentration in lateral flow and groundwater flow) was more sensitive in quick flow SS load. Parameters that control overland processes, e.g. CN2 (the curve number), OV_N (overland flow Manning's N value) and SLSUBBSN (sub–basin slope length), were found to be much more sensitive for quick flow SS load estimations.

Of the sensitive parameters, BC4 (ORGP mineralization rate) was particularly sensitive for the simulation of base flow MINP load (Fig. ~~6c~~7c). RCN (nitrogen concentration in rainfall) related specifically to the dynamics of the base flow NO_3–N load and NPERCO (nitrogen percolation coefficient) significantly affected quick flow NO_3–N load (Fig. ~~6d~~7d). Parameter CH_ONCO (channel ORGN concentration) similarly affected both flow components of ORGN load (Fig. ~~6e~~7e) and SOL_CBN (organic carbon content) was most sensitive for the simulations of quick flow ORGN and NH_4–N loads. Parameter BC1 (nitrification rate in reach) was particularly sensitive for the simulation of base flow NH_4–N load (Fig. ~~6f~~7f).

4 Discussion

~~4.1 Temporal dynamics of model performance~~

This study examined temporal dynamics of model performance and parameter sensitivity in a SWAT model application that was configured for a small, relatively steep and lower order stream catchment in New Zealand. This country faces increasing pressures on freshwater resources (Parliamentary Commissioner for the Environment, 2013) and models such as SWAT potentially offer valuable tools to inform management of water resources although, to date, the SWAT model has received limited consideration in New Zealand (Cao et al., 2006). Model evaluation on the basis of the data collected during an extended monitoring programme enabled a detailed examination of how model performance varied during different flow regimes. It also permitted error in daily mean estimates of

contaminant loads to be quantified with relative precision, allowing assessment of the ability of SWAT model to simulate contaminant loads during storm events when lower–order streams typically exhibit considerable sub–daily variability in both discharge and contaminant concentrations (Zhang et al., 2010). Separating discharge and loads of sediments and nutrients into those associated with base flow and quick flow for separate OAT sensitivity analyses provided important insights into the varying dependency of parameter sensitivity on hydrologic conditions.

4.1 Temporal dynamics of model performance

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. 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).

~~The poor fit between simulated daily mean TP concentrations and monthly instantaneous measurements may partly reflect a mismatch between the dominant processes affecting phosphorus cycling in the stream and those represented in SWAT. 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 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).~~

~~Overestimation~~ of TN concentrations prior to 1996 ~~(PBIAS = 26.7%)~~ reflects ~~the fact~~ higher ~~that~~ NO_3–N concentrations in groundwater ~~were likely higher~~ during the calibration period (~~PBIAS = 0.05%;~~ 2004–2008) due to the

Commented [MW55]: Reviewer #2, Comment #21

Commented [MW56]: Reviewer #2, Comment #19

wastewater irrigation operation~~.s,~~ Nitrate concentrations appeared to ~~and had~~ reach~~ed~~ a new quasi–steady state as ~~between~~ wastewater loads and in–stream attenuation came into balance. ~~Our decision to deliberately select a validation period (1994–1997) during which the boundary conditions of system (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). 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 to changes to sources in modified catchments (Bain et al., 2012).~~ 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.

Commented [MW57]: Reviewer #3, Comment #20

The poor fit between simulated daily mean TP concentrations and monthly instantaneous measurements may partly reflect a mismatch between the dominant processes affecting phosphorus cycling in the stream and those represented in SWAT. 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).

Commented [MW58]: Reviewer #2, Comment #6 (iii)

Our finding that measured Q–weighted mean concentrations (C_{QWM}) of TP and SS during storm events (2010–2012) were greatly underestimated relative to simulated daily mean TP ~~(PBIAS = 69.4%)~~ and SS ~~(PBIAS = 43.9%)~~ concentrations has important implications for studies that examine effects of altered flow regimes on contaminant transport. For example, studies which

simulate scenarios comprising more frequent large rainfall events (associated with climate change predictions for many regions; IPCC, 2013) may considerably underestimate projected future loads of SS and associated particulate nutrients if only base flow water quality measurements (i.e. those predominantly collected during 'state of environment' monitoring) are used for calibration/validation (see Radcliffe et al., 2009 for a discussion of this issue in relation to phosphorus). This is also reflected by the two model performance statistics relating to validation of modelled SS concentrations using monthly grab samples (predominantly base flow; 'very good') and C_{QWM} estimated during storm sampling ('unsatisfactory') based on R^2 and NSE values. ~~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. 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.~~

4.2 Key uncertainties

~~Lindenschmidt et al. (2007)~~ 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. 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.

Commented [MW59]: Reviewer #3, Comment #15 (iv)

Commented [MW60]: Reviewer #1, Comment #11
Reviewer #2, Comment #16
Reviewer #3, Comment #17 (iii)

Commented [MW61]: Reviewer #3, Comment #8, #16, and #21

4.~~2~~ 3 Temporal dynamics of parameter sensitivity

To date, studies of temporal variability of parameters have focused on hydrological parameters, rather than on water quality parameters. 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. Defining separate contaminant concentrations in base flow and quick flow enabled us to examine how the sensitivity of water quality parameters varied depending on hydrologic conditions.

Commented [MW62]: Reviewer #3, Comment #12

In a study of a lowland catchment (481 km^2), Guse et al. (2014) found that three groundwater parameters, RCHRG_DP (aquifer percolation coefficient), GW_DELAY (groundwater delay) and ALPHA_BF (base flow alpha factor) were highly sensitive in relation to simulating discharge during quick flow, while ESCO (soil evaporation compensation factor) was most sensitive during base flow. This is counter to the findings of this study for which the base–flow discharge

simulation was sensitive to RCHRG_DP and ALPHA_BF. This result may reflect that, relative to our study catchment, the catchment studied by Guse et al. (2014) had moderate precipitation (884 mm y^{-1}) with less forest cover and flatter topography. Although the GW_DELAY parameter reflects the time lag that it takes water in the soil water to enter the shallow aquifers, its lack of sensitivity under both base flow and quick flow conditions in this study is a reflection of higher water infiltration rates and steeper slopes. The ESCO parameter controls the upwards movement of water from lower soil layers to meet evaporative demand (Neitsch et al., 2011). Its lack of sensitivity in our study may reflect relatively high and seasonally–consistent rainfall (1500 mm y^{-1}), in addition to extensive forest cover in the Puarenga Stream catchment, which reduces soil evaporative demand by shading. Soil texture is also likely a contributor to this result. The predominant soil horizon type in the Puarenga Stream catchment was A, indicating high macroporosity which promotes high water infiltration rate and inhibits upward transport of water by capillary action (Neitsch et al., 2011). The variability in the sensitivity of the parameter SURLAG (surface runoff lag coefficient) between this study (relatively insensitive) and that of Cibin et al. (2010; relatively sensitive) likely reflects differences in catchment size. The Puarenga Stream catchment (77 km^2) is much smaller than the study catchment (St Joseph River; 2800 km^2) of Cibin et al. (2010) and, consequently, distances to the main channel are much shorter, with less potential for attenuation of surface runoff in off–channel storage sites. The curve number (CN2) parameter was ~~not~~ found to be ~~sensitive~~ insensitive in both this study and Shen et al. (2012), because surface runoff was simulated based on the Green and Ampt method (1911) requiring the hourly rainfall inputs, rather than the curve number equation which is an empirical model. By contrast, the most sensitive parameters in our study are those that determine the extent of lateral flow, an important contributor to streamflow in the catchment, due to a general lack of ground cover under plantation trees and formation of gully networks on steep terrain.

Commented [MW63]: Reviewer #3, Comment #23

Parameters that control surface water transport processes (e.g. LAT_TIME and SLSOIL) were found to be much more sensitive for base flow SS load estimation than parameters that control groundwater processes (e.g. ALPHA_BF and RCHRG_DP), reflecting the importance of surface flow

processes for sediment transport. Sensitive parameters for quick flow SS load estimation related to overland flow processes (e.g. OV_N and SLSUBBSN), thus reflecting the fact that sediment transport is largely dependent on rainfall–driven processes, as is typical of steep and lower–order catchments. Modelled base flow NO3–N loads were most sensitive to the nitrogen concentration in rainfall (RCN) because of rainfall as a predominant contributor to recharging base flow. The nitrogen percolation coefficient (NPERCO) was more influential for quick flow NO3–N load estimation, probably indicating that the quick flow NO3–N load is more influenced by the mobilisation of concentrated nitrogen sources associated with agriculture or treated wastewater distribution. High sensitivity of the organic carbon content (SOL_CBN) for quick flow ORGN load estimates likely reflects mobilisation of N associated with organic material following rainfall. The finding that base flow NH4–N load was more sensitive to nitrification rate in reach (BC1) likely reflects that base flow provides more favourable conditions to complete this oxidation reaction, as NH4–N is less readily leached and transported. Similarly, the ORGP mineralization rate (BC4) strongly influenced base flow MINP load estimation, reflecting that base flow phosphorus transport is relatively more influenced by cycling from channel bed stores, whereas quick flow phosphorus transport predominantly reflects the transport of phosphorus that originated from sources distant from the channel.

5 Conclusions

The performance of a SWAT model was quantified for different hydrologic conditions in a small catchment with mixed land use. Discharge–weighted mean concentrations of TP and SS measured during storm events were greatly underestimated by SWAT, highlighting the potential for uncertainty to be greatly underestimated in catchment model applications that are validated using a sample of contaminant load measurements that is over–represented by measurements made during base flow conditions. Accurate simulation of nitrogen concentrations was constrained by the non–steady state of groundwater nitrogen concentrations due to historic variability in anthropogenic nitrogen applications to land. The

sensitivity of many parameters varied depending on the relative dominance of base flow and quick flow, while curve number, soil evaporation compensation factor, surface runoff lag coefficient, and groundwater delay were largely invariant to the two flow regimes. Parameters relating to main channel processes were more sensitive when estimating variables (particularly Q and SS) during base flow, while those relating to overland processes were more sensitive for simulating variables associated with quick flow. Temporal dynamics of both parameter sensitivity and model performance due to dependence on hydrologic conditions should be considered in further model applications. Monitoring programmes which collect high–frequency and event–based data have an important role in supporting the robust calibration and validation of SWAT model applications. 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.

Commented [MW64]: Reviewer #1, Comment #13

Commented [MW65]: Reviewer #2, Comment #20

Acknowledgements

This study was funded by the Bay of Plenty Regional Council and the Ministry of Business, Innovation and Employment (Outcome Based Investment in Lake Biodiversity Restoration UOWX0505). We thank the Bay of Plenty Regional Council (BoPRC), Rotorua District Council (RDC) and Timberlands Limited for assistance with data collection. In particular, we thank Alison Lowe (RDC), Alastair MacCormick (BoPRC), Craig Putt (BoPRC) and Ian Hinton (Timberlands Limited). Theodore Kpodonu (University of Waikato) is thanked for assisting with manuscript preparation.

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.

Abbaspour, K.C., Johnson, C.A., and van Genuchten, M.T.H.: Estimating uncertain flow and transport parameters using a sequential uncertainty fitting procedure, Vadose Zone J., 3, 1340–1352, doi: 10.2136/vzj2004.1340, 2004.

Abbaspour, K.C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., and Srinivasan, R.: Modelling hydrology and water quality in the pre–alpine/alpine Thur watershed using SWAT, J. Hydrol., 333, 413–430, doi: 10.1016/j.jhydrol.2006.09.014, 2007.

Abell, J.M. and Hamilton, D.P.: Bioavailability of phosphorus transported during storm flow to a eutrophic polymictic lake, New Zeal. J. Mar. Fresh., 47, 481–489, doi: 10.1080/00288330.2013.792851, 2013.

Abell, J.M., Hamilton, D.P., and Rutherford, J.C.: Quantifying temporal and spatial variations in sediment, nitrogen and phosphorus transport in stream inflows to a large eutrophic lake, Environ. Sci.: Processes Impacts, 15, 1137–1152, doi: 10.1039/c3em00083d, 2013.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., and Williams, J.R.: Large area hydrologic modeling and assessment Part I: Model development, J. Am. Water Resour. As., 34, 73–89, doi: 10.1111/j.1752-1688.1998.tb05961.x, 1998.

Bain, D.J., Green, M.B., Campbell, J.L., Chamblee, J.F., Chaoka, S., Fraterrigo, J.M., Kaushal, S.S., Martin, S.L., Jordan, T.E., and Parolari, A.J.: Legacy effects in material flux: structural catchment changes predate long–term studies, BioScience, 62, 575–584, doi: 10.1525/bio.2012.62.6.8, 2012.

Bi, H.Q., Long, Y.S., Turner, J., Lei, Y.C., Snowdon, P., Li, Y., Harper, R., Zerihun, A., and Ximenes, F.: Additive prediction of aboveground biomass for *Pinus radiata* (D. Don) plantations, Forest Ecol. Manag., 259, 2301–2314, doi: 10.1016/j.foreco.2010.03.003, 2010.

Commented [MW66]: Reviewer #3, Comment #7 (iii)

Bieroza, M.Z., Heathwaite, A.L., Mullinger, N.J., and Keenan, P.O.: Understanding nutrient biogeochemistry in agricultural catchments: the challenge of appropriate monitoring frequencies, Environ. Sci.: Processes Impacts, 16, 1676–1691, doi: 10.1039/c4em00100a, 2014.

Boyle, D.P., Gupta, H.V., and Sorooshian, S.: Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, Water Resour. Res., 36, 3663–3674, doi: 10.1029/2000WR900207, 2000.

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, J. Hydrol., 476, 410–425, doi: 10.1016/j.jhydrol.2012.11.012, 2013.

Cao, W., Bowden, W.B., Davie, T., and Fenemor, A.: Multi–variable and multi–site calibration and validation of SWAT in a large mountainous catchment with high spatial variability, Hydrol. Process., 20, 1057–1073, doi: 10.1002/hyp.5933, 2006.

Chiwa, M., Ide, J., Maruno, R., Higashi, N., and Otsuki, K.: Effects of storm flow samplings on the evaluation of inorganic nitrogen and sulfate budgets in a small forested watershed, Hydrol. Process., 24, 631–640, doi: 10.1002/hyp.7557, 2010.

Choi, H.T. and Beven, K.J.: Multi–period and multi–criteria model conditioning to reduce prediction uncertainty in an application of TOPMODEL within the GLUE framework, J. Hydrol. (NZ), 332, 316–336, doi: 10.1016/j.jhydrol.2006.07.012, 2007.

Chow, V.T.: Open–channel hydraulics, Blackburn Press, Caldwell, New Jersey, 2008.

Cibin, R., Sudheer, K.P., and Chaubey, I.: Sensitivity and identifiability of stream flow generation parameters of the SWAT model, Hydrol. Process., 24, 1133–1148, doi: 10.1002/hyp.7568, 2010.

Conan, C., Bouraoui, F., Turpin, N., de Marsily, G., and Bidglio, G.: Modelling flow and nitrate fate at catchment scale in Britany (France), J. Environ. Qual., 32, 2026–2032, doi:10.2134/jeq2003.2026, 2003.

Dairying Research Corporation, AgResearch, Fert Research: Fertilizer use on New Zealand Dairy Farms, In New Zealand Fertiliser Manufacturers'

Research Association, Roberts, A.H.C. and Morton, J.D. (eds), Auckland, New Zealand, 36, 1999.

Eckhardt, K. and Arnold, J.G.: Automatic calibration of a distributed catchment model, J. Hydrol., 251, 103–109, 2001.

Ekanayake, J. and Davie, T.: The SWAT model applied to simulating nitrogen fluxes in the Motueka River catchment, Landcare Research ICM Report 2004–05/04, Landcare Research, Lincoln, New Zealand, 18, 2005.

Environment Bay of Plenty: Historical data summary, Report prepared for Bay of Plenty Regional Council, New Zealand, 522. 2007.

Fert Research: Fertilizer Use on New Zealand Sheep and Beef Farms, In New Zealand Fertiliser Manufacturers' Research Association, Balance, J.M. and Ravensdown, A.R. (eds), Newmarket, Auckland, New Zealand, 52, 2009.

Gassman, P.W., Reyes, M.R., Green, C.H., and Arnold, J.G.: The Soil and Water Assessment Tool: Historical development, applications, and future research directions, T. ASABE, 50, 1211–1250, 2007.

Glover, R.B.: Rotorua Chemical Monitoring to June 1993, GNS Client Report prepared for Bay of Plenty Regional Council, #722305.14, Bay of Plenty Regional Council, New Zealand, 38, 1993.

Green, W.H. and Ampt, G.A.: Studies on soil physics, part I – the flow of air and water through soils, J. Agr. Sci., 4, 1–24, doi: 10.1017/S0021859600001441, 1911.

Gupta, H.V., Sorooshian, S., and Yapo, P.O.: Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, J. Hydrol. Eng., 4, 135–143, doi: 10.1061/(ASCE)1084–0699(1999)4:2(135), 1999.

Guse, B., Reusser, D.E., and Fohrer, N.: How to improve the representation of hydrological processes in SWAT for a lowland catchment–temporal analysis of parameter sensitivity and model performance, Hydrol. Process., 28, 2651–2670, doi: 10.1002/hyp.9777, 2014.

Hall, G.M.J., Wiser, S.K., Allen, R.B., Beets, P.N., and Goulding, C.J.: Strategies to estimate national forest carbon stocks from inventory data: the 1990 New Zealand baseline, Glob. Change Biol., 7, 389–403, doi: 10.1046/j.1365–2486.2001.00419.x, 2001.

Hopmans, P. and Elms, S.R.: Changes in total carbon and nutrients in soil profiles and accumulation in biomass after a 30–year rotation of *Pinus radiata* on podzolized sands: Impacts of intensive harvesting on soil resources, Forest Ecol. Manag., 258, 2183–2193, doi: 10.1016/j.foreco.2009.02.010, 2009.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Stocker, T.F., Qin, D., Plattner, G.K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P.M. (eds), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535, 2013.

Jowett, I.: Instream habitat and minimum flow requirements for the Waipa Stream, Ian Jowett Consulting Client report: IJ0703, Report prepared for Rotorua District Council, Rotorua, New Zealand, 31, 2008.

Kirschbaum, M.U.F. and Watt, M.S.: Use of a process–based model to describe spatial variation in *Pinus radiata* productivity in New Zealand, Forest Ecol. Manag., 262, 1008–1019, doi: 10.1016/j.foreco.2011.05.036, 2011.

Krause, P., Boyle, D.P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Advances in Geosciences, 5, 89–97, 2005.

Kusabs, I. and Shaw, W.: An ecological overview of the Puarenga Stream with particular emphasis on cultural values: prepared for Rotorua District Council and Environment Bay of Plenty, Rotorua, New Zealand, 42, 2008.

Lane, L.J.: Chapter 19: Transmission Losses, In Soil Conservation Service, National engineering handbook, section 4: hydrology, U.S. Government Printing Office, Washington, D.C., 19-1–19-21, 1983.

Ledgard, S. and Thorrold, B.: Nitrogen Fertilizer Use on Waikato Dairy Farms, AgResearch and Dexcel, New Zealand, 5, 1998.

Lim, K.J., Engel, B.A., Tang, Z., Choi, J., Kim, K., Muthukrishnan, S., and Tripathy, D.: Automated Web GIS–based Hydrograph Analysis Tool, WHAT, J. Am. Water Resour. As., 41, 1407–1416, doi: 10.1111/j.1752-1688.2005.tb03808.x, 2005.

Lindenschmidt, K., Fleischbein, K., and Baborowski, M.: Structural uncertainty in a river water quality modelling system, Ecol. Model., 204, 289–300, doi: 10.1016/j.ecolmodel.2007.01.004, 2007.

Lowe, A., Gielen, G., Bainbridge, A., and Jones, K.: The Rotorua Land Treatment Systems after 16 years, In New Zealand Land Treatment Collective–Proceedings for 2007 Annual Conference, Rotorua, 14–16 March 2007, 66–73, 2007.

Mahon, W.A.J.: The Rotorua geothermal field: technical report of the Geothermal Monitoring Programme, 1982–1985, Ministry of Energy, Oil and Gas Division, Wellington, New Zealand, 1985.

Marino, S., Hogue, I.B., Ray, C.J., and Kirschner, D.E.: A methodology for performing global uncertainty and sensitivity analysis in systems biology, J. Theor. Biol., 254, 178–196, doi: 10.1016/j.jtbi.2008.04.011, 2008.

McKenzie, B.A., Kemp, P.D., Moot, D.J., Matthew, C., and Lucas, R.J.: Environmental effects on plant growth and development, In New Zealand Pasture and Crop Science, White, J.G.H. and Hodgson, J. (eds), Oxford University Press: Auckland, New Zealand, 29–44, 1999.

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.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., and Veith, T.L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, T. ASAE, 50, 885–900, 2007.

Morris, M.D.: Factorial sampling plans for preliminary computational experiments, Technometrics, 33, 161–174, 1991.

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., and Williams, J.R.: Soil and Water Assessment Tool Theoretical Documentation Version 2009, Texas Water Resources Institute Technical Report No. 406, Texas A&M University System, College Station, Texas, 647, 2011.

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.

Nielsen, A., Trolle, D., Me, W., Luo, L.C., Han, B.P., Liu, Z.W., Olesen, J.E., and Jeppesen, E.: Assessing ways to combat eutrophication in a Chinese drinking water reservoir using SWAT, Mar. Freshwater Res., 64, 475–492, doi: 10.1071/MF12106, 2013.

Commented [MW67]: Reviewer #1, Comment #12

Commented [MW68]: Reviewer #3, Comment #8

Paku, L.K.: The use of carbon–13 to trace the migration of treated wastewater and the chemical composition in a forest environment, Master Thesis, Science in Chemistry, the University of Waikato, Hamilton, New Zealand, 92, 2001.

Parliamentary Commissioner for the Environment: Water Quality in New Zealand: Land Use and Nutrient Pollution, New Zealand, 82, 2013.

Pfannerstill, M., Guse, B., and Fohrer, N.: Smart low flow signature metrics for an improved overall performace evaluation of hydrological models, J. Hydrol., 510, 447–458, 2014.

Radcliffe, D.E., Lin, Z., Risse, L.M., Romeis, J.J., and Jackson, C.R.: Modeling phosphorus in the Lake Allatoona watershed using SWAT: I. Developing phosphorus parameter values, J. Environ. Qual., 38, 111–120, doi:10.2134/jeq2007.0110, 2009.

Reusser, D.E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrol. Earth. Syst. Sc., 13, 999–1018, doi:10.5194/hess-13-999-2009, 2009.

Reusser, D.E. and Zehe, E.: Inferring model structural deficits by analysing temporal dynamics of model performance and parameter sensitivity, Water Resour. Res., 47, W07550, 15pp, doi: 10.1029/2010WR009946, 2011.

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

Rimmer, A. and Hartmann, A.: Optimal hydrograph separation filter to evaluate transport routines of hydrological models, J. Hydrol., 514, 249–257, doi: 10.1016/j.jhydrol.2014.04.033, 2014.

Rotorua District Council, Rotorua Wastewater Treatment Plant, Rotorua, New Zealand, 22, 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.

Sangrey, D.A., Harrop–Williams, K.O., and Klaiber, J.A.: Predicting ground–water response to precipitation, J. Geotech. Eng., 110, 957–975, doi: 10.1061/(ASCE)0733–9410(1984)110:7(957), 1984.

Commented [MW69]: Reviewer #3, Comment #7 (iii)

Commented [MW70]: Reviewer #2, Comment #21

Schuol, J., Abbaspour, K.C., Yang, H., and Srinivasan, R.: Modeling blue and green water availability in Africa, Water Resour. Res., 44, W07406, 18 pp, doi: 10.1029/2007WR006609, 2008.

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

Shen, Z.Y., Chen, L., and Chen, T.: Analysis of parameter uncertainty in hydrological and sediment modeling using GLUE method: a case study of SWAT model applied to Three Gorges Reservoir Region, China, Hydrol. Earth. Syst. Sc., 16, 121–132, doi: 10.5194/hess–16–121–2012, 2012.

Sloan, P.G. and Moore, I.D.: Modelling subsurface stormflow on steeply sloping forested watersheds, Water Resour. Res., 20, 1815–1822, doi: 10.1029/WR020i012p01815, 1984.

Statistics New Zealand: Fertiliser use in New Zealand, Statistics New Zealand, New Zealand, 13, 2006.

van Griensven, A., Meixner, T., Grunwald, S., Bishop, T., Diluzio, M., and Srinivasan, R.: A global sensitivity analysis tool for the parameters of multi–variable catchment models, J. Hydrol., 324, 10–23, doi: 10.1016/j.jhydrol.2005.09.008, 2006.

Watt, M.S., Clinton, P.W., Coker, G., Davis, M.R., Simcock, R., Parfitt, R.L., and Dando, J.: Modelling the influence of environment and stand characteristics on basic density and modulus of elasticity for young *Pinus radiata* and *Cupressus lusitanica*, Forest Ecol. Manag., 255, 1023–1033, doi: 10.1016/j.foreco.2007.09.086, 2008.

White, K.L. and Chaubey, I.: Sensitivity analysis, calibration, and validations for a multisite and multivariable SWAT model, J. Am. Water Resour. As., 41, 1077–1089, doi: 10.1111/j.1752-1688.2005.tb03786.x, 2005.

White, M.J., Storm, D.E., Mittelstet, A., Busteed, P.R., Haggard, B.E., and Rossi, C.: Development and testing of an in–stream phosphorus cycling model for the Soil and Water Assessment Tool, J. Environ. Qual., 43, 215–223, doi: 10.2134/jeq2011.0348, 2014.

White, P.A., Cameron, S.G., Kilgour, G., Mroczek, E., Bignall, G., Daughney, C., and Reeves, R.R.: Review of groundwater in Lake Rotorua catchment, Prepared for Environment Bay of Plenty, Institute of Geological &

Commented [MW71]: Reviewer #3, Comment #8

Nuclear Sciences Client Report 2004/130, Whakatane, New Zealand, 245, 2004.

Whitehead, D., Kelliher, F.M., Lane, P.M., and Pollock, D.S.: Seasonal partitioning of evaporation between trees and understorey in a widely spaced *Pinus radiata* stand, J. Appl. Ecol., 31, 528–542, 1994.

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.

Ximenes, F.A., Gardner, W.D., and Kathuria, A.: Proportion of above–ground biomass in commercial logs and residues following the harvest of five commercial forest species in Australia, Forest Ecol. Manag., 256, 335–346, doi: 10.1016/j.foreco.2008.04.037, 2008.

Yilmaz, K.K., Gupta, H.V., and Wagener, T.: A process–based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, 18 pp, doi: 10.1029/2007WR006716, 2008.

Zhang, H., Huang, G.H., Wang, D.L., and Zhang, X.D.: Multi–period calibration of a semi–distributed hydrological model based on hydroclimatic clustering, Adv. Water Resour., 34, 1292–1303, 2011.

Zhang, Z., Tao, F., Shi, P., Xu, W., Sun, Y., Fukushima, T., and Onda, Y.: Characterizing the flush of stream chemical runoff from forested watersheds, Hydrol. Process., 24, 2960–2970, doi: 10.1002/hyp.7717, 2010.

Commented [MW72]: Reviewer #1, Comment #4

Table 1. Description of data used to configure and calibrate the SWAT model.

Data	Application	Data description and configuration details	Source
Digital elevation model (DEM) & digitized stream network	Sub–basin delineation (Fig. 1b)	25 m resolution. Used to define five slope classes: 0–4%, 4–10%, 10–17%, 17–26% and >26%.	Bay of Plenty Regional Council (BoPRC)
Stream discharge and water quality measurements	Calibration (2004–2008) and validation[1] (1994–1997; 2010–2012)	FRI: 15–min stream discharge data were aggregated as daily mean values (1994–1997; 2004–2008), monthly grab samples for determination of ~~instantaneous~~ suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) concentrations (1994–1997; 2004–2008), high–frequency event–based samples for concentrations of SS (nine events), TP and TN (both 14 events) at 1–2 h frequency (2010–2012).	BoPRC; Abell et al., 2013
Spring discharge~~,~~ and nutrient loads~~, and water abstraction volumes~~	Point source (Fig. 1b) ~~and water use~~	Constant daily discharge and nutrient concentrations assigned to two cold–water springs (Waipa Spring and Hemo Spring) and one geothermal spring. ~~based on spot measurements. Constant nutrient concentrations assigned to Waipa Spring and Hemo Spring and the geothermal spring based on samples collected between August 1984 and June 2004. Monthly water abstraction assigned to two cold–water springs.~~	~~Kusabs and Shaw, 2008;~~ White et al., 2004; Proffit, 2009 (Unpublished Site Visit Report); Paku, 2001; Mahon, 1985; Glover, 1993; ~~Jowett, 2008;~~ Rotorua District Council (pers. comm.)
Water abstraction volumes	Water use	Monthly water abstraction assigned to two cold–water springs.	Kusabs and Shaw, 2008; Jowett, 2008
Land use	HRU definition	25 m resolution, 10 basic land–cover categories. Some particular land–cover parameters were prior–estimated	New Zealand Land Cover Database Version 2; BoPRC

[1] 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.

Commented [MW73]: Reviewer #2, Comment #8 (ii)

Commented [MW74]: Reviewer #2, Comment #8 (i)

Commented [MW75]: Reviewer #2, Comment #8 (iii)

Commented [MW76]: Reviewer #2, Comment #8 (iii)

		(Table 2).	
Soil characteristics	HRU definition	~~Properties of~~ 22 soil types. Properties were quantified based on measurements (if available) or estimated using regression analysis to estimate properties for unmeasured functional horizons. ~~were determined using the key physical properties and the characteristics of functional horizons provided by soil map~~	New Zealand Land Resource Inventory & digital soil map (available at http://smap.landcareresearch.co.nz)
Meteorological data	Meteorological forcing	Daily maximum and minimum temperature, daily mean relative humidity, daily global solar radiation, daily (9 am) surface wind speed and hourly precipitation.	Rotorua Airport Automatic Weather Station, National Climate Database ~~National Climatic Data Centre~~ (available at http://cliflo.niwa.co.nz/); Kaituna rain gauge (Fig. 1a)
Agricultural management practices	Agricultural management schedules	~~Farm–specific stocking density, fertilizer application rates and farming practices (1993–2012). Simulated applications of urea (twice in winter/spring; four times in summer/autumn) and di–ammonium phosphate (once or twice in spring/autumn). Application of manure–associated nutrients to paddocks was simulated as a function of stock numbers and literature values for the average N and P content of excreta.~~Stock density	Statistics New Zealand, 2006; ~~Fert Research, 2009;~~ Ledgard and Thorrold, 1998~~; Dairying Research Corporation, 1999~~
		Applications of urea and di–ammonium phosphate	Statistics New Zealand, 2006; Fert Research, 2009
		Applications of manure–associated nutrients	Dairying Research Corporation, 1999
Nutrient loading by wastewater application	Nonpoint–source from land treatment irrigation	Wastewater application rates and effluent composition (TN and TP concentration) for 16 spray blocks from 1996–2012. Each spray block was assigned an individual management schedule specifying daily application rates.	Rotorua District Council, 2006
Forest stand map and harvest dates	Forestry planting and harvesting	Planting and harvesting data for 472 ha forestry stands. Prior to 2007 we assumed stands were cleared one–year prior to the establishment year. Post 2007, harvesting	Timberlands Limited, Rotorua, New Zealand (pers. comm.)

Commented [MW77]: Reviewer #2, Comment #9 (i)

Commented [MW78]: Reviewer #2, Comment #9 (ii)

Commented [MW79]: Reviewer #2, Comment #9 (iii)

Commented [MW80]: Reviewer #2, Comment #9 (iii)

	operations	date was assigned to the first day of harvesting month.

Table 2. Prior–estimated parameter values for three dominant types of land–cover in the Puarenga Stream catchment. Values of other land use parameters were based on the default values in the SWAT database.

Land–cover type	Parameter	Definition	Value	Source
PINE (*Pinus radiata*)	HVSTI	Percentage of biomass harvested	0.65	(Ximenes et al., 2008)
	T_OPT (℃)	Optimal temperature for plant growth	15	(Kirschbaum and Watt 2011)
	T_BASE (℃)	Minimum temperature for plant growth	4	(Kirschbaum and Watt 2011)
	MAT_YRS	Number of years to reach full development	30	(Kirschbaum and Watt 2011)
	BMX_TREES (tonnes ha^{-1})	Maximum biomass for a forest	400	(Bi et al., 2010)
	GSI (m s^{-1})	Maximum stomatal conductance	0.00198	(Whitehead et al., 1994)
	BLAI (m^2 m^{-2})	Maximum leaf area index	5.2	(Watt et al., 2008)
	BP3	Proportion of P in biomass at maturity	0.000163	(Hopmans and Elms 2009)
	BN3	Proportion of N in biomass at maturity	0.00139	(Hopmans and Elms 2009)
FRSE (Evergreen forest)	HVSTI	Percentage of biomass harvested	0	–
	BMX_TREES (tonnes ha^{-1})	Maximum biomass for a forest	372	(Hall et al., 2001)
	MAT_YRS (years)	Number of years for tree to reach full development	100	–
PAST (Pastoral farm)	T_OPT (℃)	Optimal temperature for plant growth	25	(McKenzie et al., 1999)
	T_BASE (℃)	Minimum temperature for plant growth	5	(McKenzie et al., 1999)

Table 3. Summary of calibrated SWAT parameters. Discharge (Q), suspended sediment (SS) and total nitrogen (TN) parameter values were assigned using auto–calibration, while total phosphorus (TP) parameters were manually calibrated. SWAT default ranges and input file extensions are shown for each parameter.

Parameter	Definition	Unit	Default range	Calibrated value
Q ~~and SS~~				
EVRCH.bsn	Reach evaporation adjustment factor		0.5–1	0.9
~~PRF.bsn~~	~~Peak rate adjustment factor for sediment routing in the main channel~~		~~0–2~~	
~~SPCON.bsn~~	~~Linear parameter for calculating the maximum amount of sediment that can be re–entrained during channel sediment routing~~		~~0.0001–0.01~~	
~~SPEXP.bsn~~	~~Exponent parameter for calculating sediment re–entrained in channel sediment routing~~		~~1–1.5~~	
SURLAG.bsn	Surface runoff lag coefficient		0.05–24	15
ALPHA_BF.gw	Base flow alpha factor (0–1)		0.0071–0.0161	0.01
GW_DELAY.gw	Groundwater delay	d	0–500	500
GW_REVAP.gw	Groundwater "revap" coefficient		0.02–0.2	0.08
GW_SPYLD.gw	Special yield of the shallow aquifer	$m^3\ m^{-3}$	0–0.4	0.13
GWHT.gw	Initial groundwater height	m	0–25	14
GWQMN.gw	Threshold depth of water in the shallow aquifer required for return flow to occur	mm	0–5000	372
RCHRG_DP.gw	Deep aquifer percolation fraction		0–1	0.87
REVAPMN.gw	Threshold depth of water in the shallow aquifer required for "revap" to occur	mm	0–500	260
CANMX.hru	Maximum canopy storage	mm	0–100	0.6
EPCO.hru	Plant uptake compensation factor		0–1	0.34
ESCO.hru	Soil evaporation compensation factor		0–1	0.9
HRU_SLP.hru	Average slope steepness	$m\ m^{-1}$	0–0.6	0.5
~~LAT_SED.hru~~	~~Sediment concentration in lateral flow and groundwater flow~~	~~$mg\ L^{-1}$~~	~~0–5000~~	

Commented [MW81]: Reviewer #2, Comment #11
Reviewer #3, Comment #7 (vii)

Commented [MW82]: Editor, Comment #5

LAT_TTIME.hru	Lateral flow travel time	d	0–180	3
~~OV_N.hru~~	~~Manning's N value for overland flow~~		~~0.01–30~~	
RSDIN.hru	Initial residue cover	kg ha^{-1}	0–10000	1
SLSOIL.hru	Slope length for lateral subsurface flow	m	0–150	40
~~SLSUBBSN.hru~~	~~Average slope length~~	~~m~~	~~10–150~~	
~~CH_COV1.rte~~	~~Channel erodibility factor~~		~~0–0.6~~	
~~CH_COV2.rte~~	~~Channel cover factor~~		~~0–1~~	
CH_K2.rte	Effective hydraulic conductivity in the main channel alluvium	mm h^{-1}	0–500	20
CH_N2.rte	Manning's N value for the main channel		0–0.3	0.16
CH_K1.sub	Effective hydraulic conductivity in the tributary channel alluvium	mm h^{-1}	0–300	100
CH_N1.sub	Manning's N value for the tributary channel		0.01–30	20
~~CN2.mgt~~	~~Initial SCS runoff curve number for moisture condition~~		~~35–89~~	
SS				
USLE_P.mgt	USLE equation support practice factor		0–1	0.5
PRF.bsn	Peak rate adjustment factor for sediment routing in the main channel		0–2	1.9
SPCON.bsn	Linear parameter for calculating the maximum amount of sediment that can be re–entrained during channel sediment routing		0.0001–0.01	0.001
SPEXP.bsn	Exponent parameter for calculating sediment re–entrained in channel sediment routing		1–1.5	1.26
LAT_SED.hru	Sediment concentration in lateral flow and groundwater flow	mg L^{-1}	0–5000	5.7
OV_N.hru	Manning's N value for overland flow		0.01–30	28
SLSUBBSN.hru	Average slope length	m	10–150	92
CH_COV1.rte	Channel erodibility factor		0–0.6	0.17
CH_COV2.rte	Channel cover factor		0–1	0.6
TP				
P_UPDIS.bsn	Phosphorus uptake distribution parameter		0–100	0.5
PHOSKD.bsn	Phosphorus soil partitioning coefficient		100–200	174
PPERCO.bsn	Phosphorus percolation coefficient		10–17.5	14

Commented [MW83]: Reviewer #3, Comment #22

Commented [MW84]: Editor, Comment #5

PSP.bsn	Phosphorus sorption coefficient		0.01–0.7	0.5
GWSOLP.gw	Soluble phosphorus concentration in groundwater loading	mg P L^{-1}	0–1000	0.063
LAT_ORGP.gw	Organic phosphorus in the base flow	mg P L^{-1}	0–200	0.01
ERORGP.hru	Organic phosphorus enrichment ratio		0–5	2.5
CH_OPCO.rte	Organic phosphorus concentration in the channel	mg P L^{-1}	0–100	0.02
BC4.swq	Rate constant for mineralization of organic phosphorus to dissolved phosphorus in the reach at 20 ℃	d^{-1}	0.01–0.7	0.3
RS2.swq	Benthic (sediment) source rate for dissolved phosphorus in the reach at 20 ℃	mg m^{-2} d^{-1}	0.001–0.1	0.02
RS5.swq	Organic phosphorus settling rate in the reach at 20 ℃	d^{-1}	0.001–0.1	0.05
TN				
RSDCO.bsn	Residue decomposition coefficient		0.02–0.1	0.09
CDN.bsn	Denitrification exponential rate coefficient		0–3	0.3
CMN.bsn	Rate factor for humus mineralization of active organic nitrogen		0.001–0.003	0.002
N_UPDIS.bsn	Nitrogen uptake distribution parameter		0–100	0.5
NPERCO.bsn	Nitrogen percolation coefficient		0–1	0.0003
RCN.bsn	Concentration of nitrogen in rainfall	mg N L^{-1}	0–15	0.34
SDNCO.bsn	Denitrification threshold water content		0–1	0.02
HLIFE_NGW.gw	Half–life of nitrate–nitrogen in the shallow aquifer	d	0–200	195
LAT_ORGN.gw	Organic nitrogen in the base flow	mg N L^{-1}	0–200	0.055
SHALLST_N.gw	Nitrate–nitrogen concentration in the shallow aquifer	mg N L^{-1}	0–1000	1
ERORGN.hru	Organic nitrogen enrichment ratio		0–5	3
CH_ONCO.rte	Organic nitrogen concentration in the channel	mg N L^{-1}	0–100	0.01
BC1.swq	Rate constant for biological oxidation of ammonium–nitrogen to nitrite–nitrogen in the reach at 20 ℃	d^{-1}	0.1–1	1
BC2.swq	Rate constant for biological oxidation of nitrite–nitrogen to nitrate–nitrogen in the reach at 20 ℃	d^{-1}	0.2–2	0.7
BC3.swq	Rate constant for hydrolysis of organic nitrogen to ammonium–nitrogen in the reach at 20 ℃	d^{-1}	0.2–0.4	0.4

RS3.swq	Benthic (sediment) source rate for ammonium–nitrogen in the reach at 20 ℃	mg m^{-2} d^{-1}	0–1	0.2
RS4.swq	Rate coefficient for organic nitrogen settling in the reach at 20 ℃	d^{-1}	0.001–0.1	0.05

Table 4. Criteria for model performance. Note: $\boldsymbol{o_n}$ is the $\boldsymbol{n^{th}}$ observed datum, $\boldsymbol{s_n}$ is the $\boldsymbol{n^{th}}$ simulated datum, $\bar{\mathbf{o}}$ is the observed mean value, $\bar{\mathbf{s}}$ is the simulated daily mean value, and **N** is the total number of observed data. Performance rating criteria are based on Moriasi et al. (2007) for Q: discharge, SS: suspended sediment, TP: total phosphorus and TN: total nitrogen. Moriasi et al. (2007) derived these criteria based on extensive literature review and analysing the reported performance ratings for recommended model evaluation statistics.

Commented [MW85]: Reviewer #1, Comment #9

Statistic equation	Constituent	Performance ratings			
		Unsatisfactory	Satisfactory	Good	Very good
$R^2 = \frac{\{\sum_{n=1}^{N}[(s_n-\bar{s})(o_n-\bar{o})]\}^2}{\sum_{n=1}^{N}(o_n-\bar{o})^2 \times \sum_{n=1}^{N}(s_n-\bar{s})^2}$ (~~2~~3)	All	< 0.5	0.5 – 0.6	0.6 – 0.7	0.7 – 1
$NSE = 1 - \frac{\sum_{n=1}^{N}(o_n - s_n)^i}{\sum_{n=1}^{N}(o_n - \bar{o})^i} \quad i = 2$ (~~3~~4)	All	< 0.5	0.5 – 0.65	0.65 – 0.75	0.75 – 1
$\pm PBIAS\% = \frac{\sum_{n=1}^{N}(o_n-s_n)}{\sum_{n=1}^{N} o_n} \times 100$ (~~4~~5)	Q	> 25	15 – 25	10 – 15	< 10
	SS	> 55	30 – 55	15 – 30	< 15
	TP, TN	> 70	40 – 70	25 – 40	< 25

R^2: coefficient of determination

NSE: Nash–Sutcliffe efficiency

PBIAS: percent bias

Table 5. Model performance ratings for simulations of discharge (Q), concentrations of suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) ~~simulations~~. n indicates the number of measurements. Q–weighted mean concentrations were calculated using Eq. (1).

Model performance	Statistics	Q	SS	TP	TN
Calibration with instantaneous measurements (2004–2008)		n = 1439	n = 43	n = 45	n = 39
	R^2	0.77 (Very good)	0.42 (Unsatisfactory)	0.02 (Unsatisfactory)	0.08 (Unsatisfactory)
	NSE	0.73 (Good)	-0.08 (Unsatisfactory)	-1.31 (Unsatisfactory)	-0.30 (Unsatisfactory)
	±PBIAS%	7.8 (Very good)	-18.3 (Very good)	23.8 (Very good)	-0.05 (Very good)
Validation with instantaneous measurements (1994–1997)		n = 1294	n = 37	n = 37	n = 36
	R^2	0.68 (Good)	0.80 (Very good)	0.01 (Unsatisfactory)	0.01 (Unsatisfactory)
	NSE	0.62 (Satisfactory)	0.76 (Very good)	-0.97 (Unsatisfactory)	-2.67 (Unsatisfactory)
	±PBIAS%	8.8 (Very good)	-0.32 (Very good)	24.5 (Very good)	-26.7 (Good)
Validation with Q–weighted mean concentrations (2010–2012)		–	n = 12	n = 18	n = 18
	R^2	–	0.38 (Unsatisfactory)	0.06 (Unsatisfactory)	0.46 (Unsatisfactory)
	NSE	–	-0.03 (Unsatisfactory)	-4.88 (Unsatisfactory)	0.42 (Unsatisfactory)
	±PBIAS%	–	43.9 (Satisfactory)	69.4 (Satisfactory)	-0.87 (Very good)

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

Commented [MW86]: Reviewer #1, Comment #11

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					

Commented [MW87]: Reviewer #2, Comment #6 (iv)
Reviewer #3, Comment #7 (iii)

Figure captions

Figure 1. (a) Location of Puarenga Stream surface catchment in New Zealand, Kaituna rain gauge, climate station and managed land areas for which management schedules were prescribed in SWAT, and (b) location of the Puarenga Stream, major tributaries, monitoring stream–gauges, two cold–water springs and the Whakarewarewa geothermal contribution.

Figure 2. Flow chart of methods used to separate hydrograph and contaminant loads and to quantify ~~for~~ parameter sensitivities~~y analysis in sequence of each individual variable~~ for: Q (discharge), SS (suspended sediment), MINP (mineral phosphorus), ORGN (organic nitrogen), NH_4–N (ammonium–nitrogen), and NO_3–N (nitrate–nitrogen). *NSE*: Nash–Sutcliffe efficiency.

Commented [MW88]: Reviewer #2, Comment #14

Figure 3. Measurements and daily mean simulated values of discharge, suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) during calibration (a–d) and validation (e–h). Measured daily mean discharge was calculated from 15–min observations and measured concentrations of SS, TP and TN correspond to monthly grab samples.

Figure 4. Example of a storm event showing derivation of ~~hourly measurements, calculated~~ discharge (Q)–weighted daily mean concentrations (dashed horizontal line) ~~from the~~based on hourly measured ~~and simulated daily mean~~ concentrations (black dots) of suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) ~~for~~ over two days ~~during one storm event~~ (a–c). Comparisons ~~includes~~ of Q–weighted daily mean concentrations with simulated daily mean estimates of SS, TP and TN (scatter plot, d–f). ~~for 24 h periods (~~The horizontal bars show the range ~~of~~ s in hourly measurements~~)~~ during each storm event~~s~~ ~~(~~in 2010–2012~~) and simulated daily mean estimates of SS, TP and TN (d–f)~~.

Commented [MW89]: Reviewer #3, Comment #15 (i)

Figure 5. Measurements and simulations derived using the calibrated set of parameter values. Data are shown separately for base flow and quick flow. (a) Daily mean base flow and quick flow; (b) suspended sediment (SS) load; (c) total phosphorus (TP) load; (d) total nitrogen (TN) load. Vertical lines in b–d show the contaminant load in quick flow. Time series relate to calibration (2004–2008) and

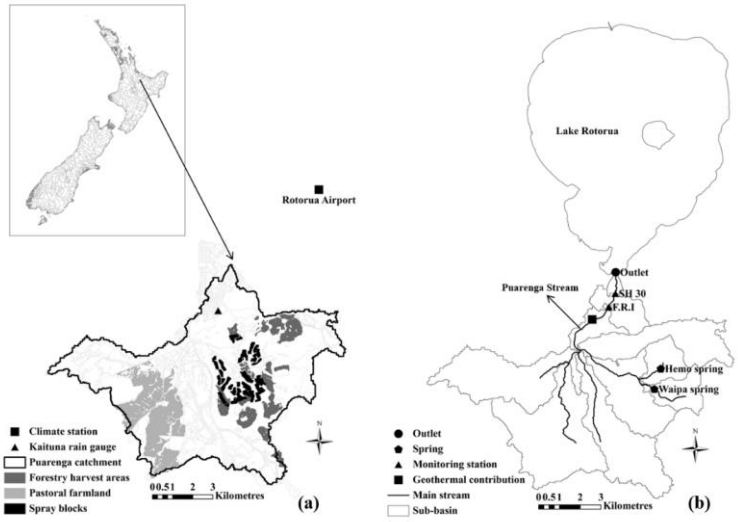
validation (1994–1997) periods (note time discontinuity). Measured instantaneous loads of SS, TP, and TN correspond to monthly grab samples.

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.

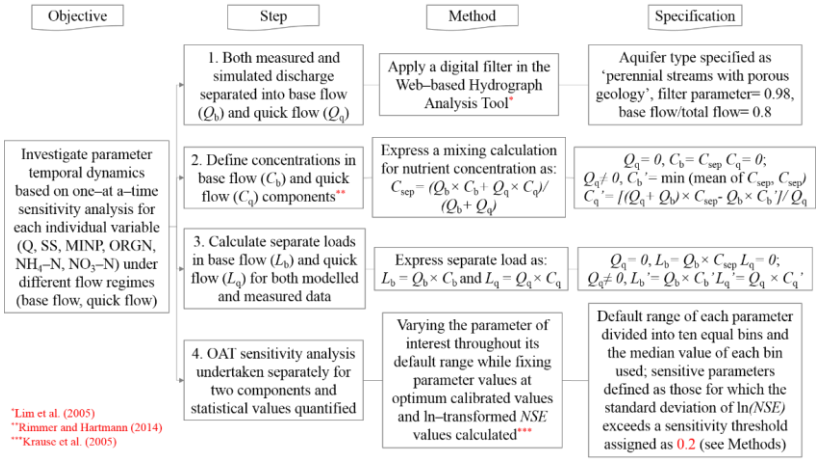
Commented [MW90]: Reviewer #3, Comment #8

Figure ~~6~~7. The standard deviation (STD) of the ~~natural log~~ln–transformed Nash–Sutcliffe efficiency (NSE) used to indicate ~~P~~parameter sensitivity ~~partitioned into for base flow and quick flow constituents~~ based on one–at a–time (OAT) sensitivity analysis for ~~each modelled and observed~~separate base and quick flow components~~simulated variable~~: (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). ~~Parameter sensitivity is quantified as the variation in standard deviation (STD) of log_{10} transformed Nash Sutcliffe efficiency (NSE).~~ A median value (0.2) derived from the STD of ln–transformed NSE was chosen ~~arbitrarily~~ as a threshold above which parameters were deemed to be 'sensitive'~~with a sensitivity threshold assigned as 0.1 (see Section 2.5)~~. Definitions of each parameter are shown in Table 3.

Commented [MW91]: Reviewer #3, Comment #18

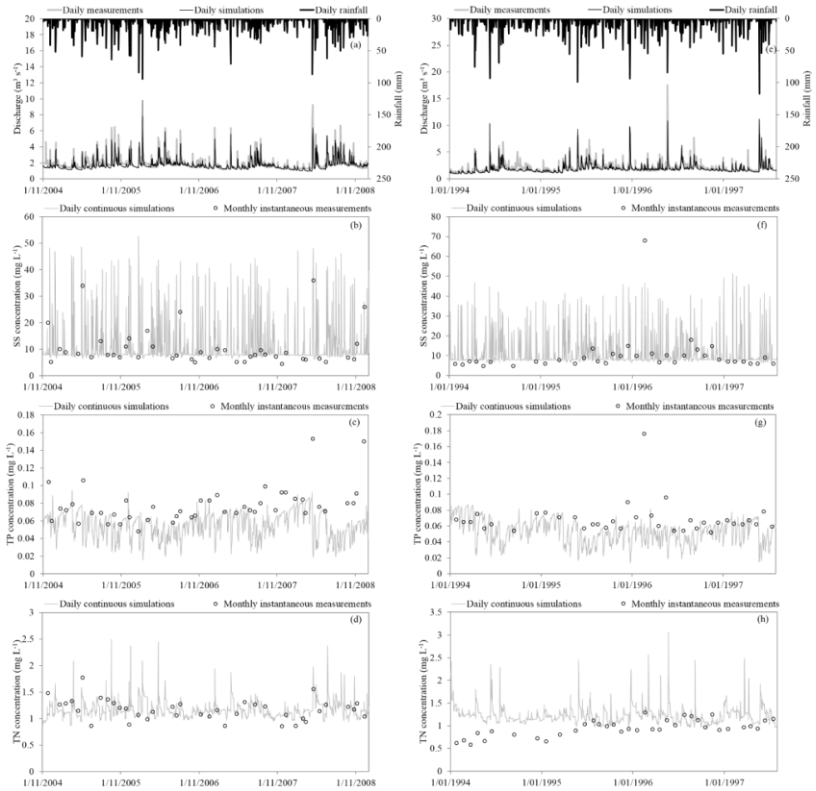


- No changes in Figure 1.

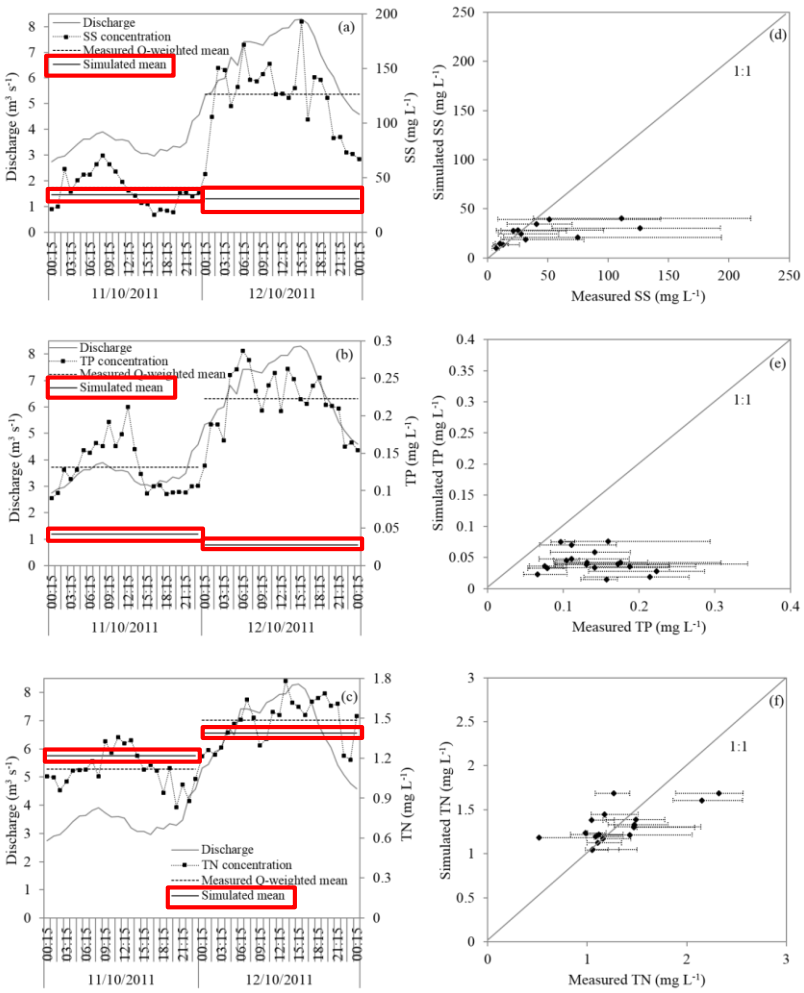


- References have been added in Figure 2 using footnotes. Specifically: "Web–based Hydrograph Analysis Tool (Lim et al. 2005)"; Define concentrations in base flow (C_b) and quick flow (C_q) components (cf. Rimmer and Hartmann, 2014); and the natural logarithm (Krause et al., 2005)".

Commented [MW92]: Reviewer #2, Comment #14

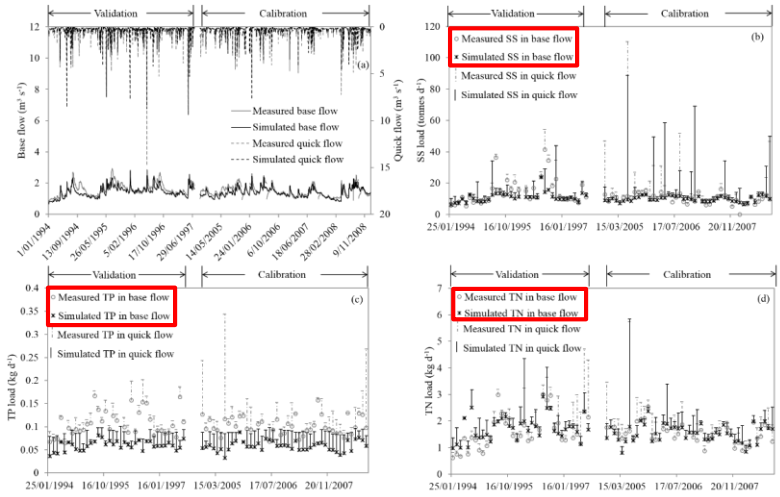


- No changes in Figure 3.



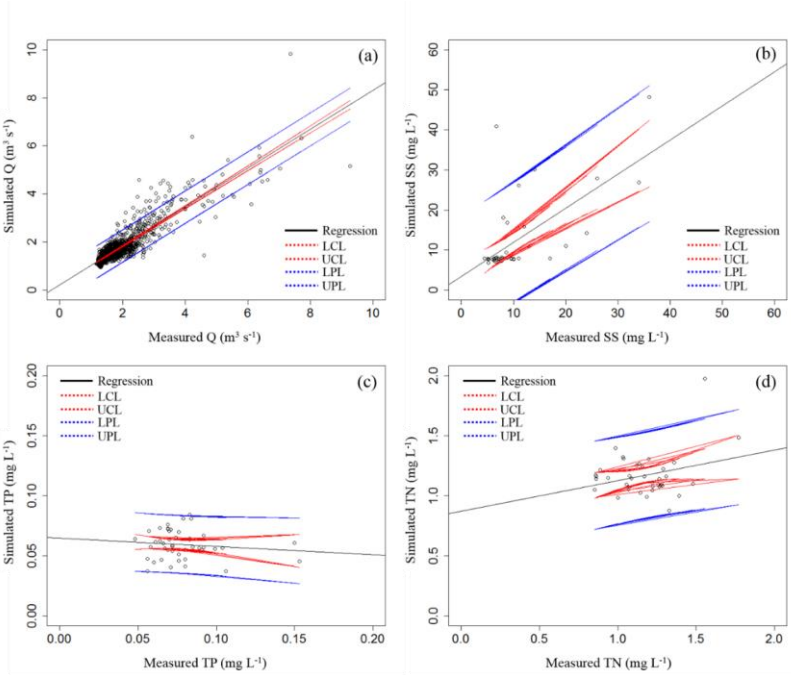
- In Figure 4 a–c, we removed the black horizontal line showing the simulated daily mean. No more changes were made as the symbols in the publisher's version appear to be clear.

Commented [MW93]: Reviewer #3, Comment #15 (i)



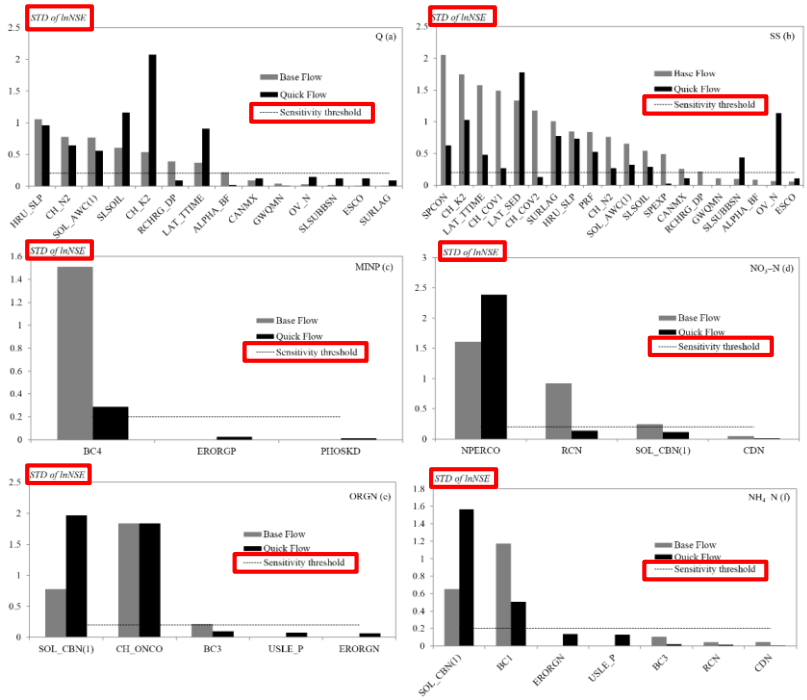
- The symbols have been made clear in Figure 5.

Commented [MW94]: Reviewer #3, Comment #17 (i)



- Figure 6 has been added to show model uncertainties for simulations of discharge (Q) and suspended sediment (SS), total phosphorus (TP) and total nitrogen (TN) concentrations.

Commented [MW95]: Reviewer #3, Comment #8



➢ The standard deviation (*STD*) of the ln–transformed NSE were used to indicate parameter sensitivity for the two flow regimes in Figure 7. The threshold value of "0.2" was then chosen based on the median value derived from the calculations of the *STD* of ln–transformed NSE".

Commented [MW96]: Reviewer #3, Comment #18