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”.

Reference:

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.