# The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model

Hollaway, M.J.[1], Beven, K.J.[1],Benskin, C.McW.H.[1], Collins, A.L.[2], Evans, R.[3], Falloon, P.D.[4], Forber, K.J.[1], Hiscock, K.M.[5], Kahana, R.[4], Macleod, C.J.A.[6], Ockenden, M.C.[1], Villamizar, M.L.[7], Wearing, C.[1], Withers, P.J.A.[8], Zhou, J.G.[9], Barber, N.J.[10] Haygarth, P.M.[1]

[1] Lancaster Environment Centre, Lancaster University, Bailrigg, Lancaster LA1 4YQ, England, UK

[2] Rothamsted Research North Wyke, Okehampton, Devon EX20 2SB, England, UK

[3] Global Sustainability Institute, Anglia Ruskin University, Cambridge CB1 1PT, England, UK

[4] Met Office Hadley Centre, Exeter, Devon EX1 3PB, England, UK

[5] School of Environmental Sciences, Norwich Research Park, University of East Anglia, Norwich NR4 7TJ, England, UK

[6] James Hutton Institute, Aberdeen AB15 8QH, Scotland, UK

[7] School of Engineering, Liverpool University, L69 3GQ, England, UK

[8] School of Environment, Natural Resources and Geography, Bangor University, Bangor, Gwynedd LL57 2UW, Wales, UK

[9] School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester M1 5GD, England, UK

[10] Geography Department, Durham University, Durham, DH1 3LE, England, UK

Corresponding author: Michael Hollaway (m.hollaway@lancaster.ac.uk)

**Abstract**

There is a need to model and predict the transfer of phosphorus (P) from land to water, but this is challenging because of the large number of complex physical and biogeochemical processes involved. This study presents, for the first time, a 'limits of acceptability' approach of the Generalized Likelihood Uncertainty Estimation (GLUE) framework to the Soil and Water Assessment Tool (SWAT), in an application to a water quality problem in the Newby Beck Catchment (12.5km$^2$), Cumbria, United Kingdom (UK). Using high frequency outlet data (discharge and P), individual evaluation criteria (limits of acceptability) were assigned to observed discharge and P loads for all evaluation time steps, identifying where the model was performing well/poorly and to infer which processes required improvement in the model structure. Initial limits of acceptability were required to be relaxed by a substantial amount (by factors of between 5.3 and 6.72 on a normalized scale depending on the evaluation criteria used) in order to gain a set of behavioral simulations (1001 and 1016, respectively out of 5,000,000). Of the 39 model parameters tested, the representation of subsurface processes and associated parameters, were consistently shown as critical to the model not meeting the evaluation criteria, irrespective of the chosen evaluation metric. It is therefore concluded that SWAT is not an appropriate model to guide P management in this catchment. This approach highlights the importance of high frequency monitoring data for setting robust model evaluation criteria. It also raises the question as to whether it is possible to have sufficient input data available to drive such models so that we can have confidence in their predictions and their ability to inform catchment management strategies to tackle the problem of diffuse pollution from agriculture.

Keywords: SWAT, GLUE, phosphorus, uncertainty analysis, River Eden, high frequency data.

## 1 Introduction

In response to water quality targets set under the Water Framework Directive (WFD) (EC 2000/60/EC European Union 2000), it is imperative that we understand the sources, mobilization and delivery of diffuse pollution from agricultural land in headwater catchments to the river network (Haygarth et al., 2005; Perks et al., 2015). In order to devise management strategies that reduce the transfer of macronutrients (e.g. phosphorus (P) and nitrogen (N)) to river networks (McGonigle et al., 2014), models are essential tools in predicting how catchments may respond to key pressures in the present and into an uncertain future. Under climate change, winters are expected to become wetter and warmer, whilst summers are predicted to be hotter and drier in the United Kingdom (UK; Jones et al., 2010). Coupled with extended periods of drought, and an increase in extreme precipitation events for much of the UK (Kendon et al., 2014), these changes are likely to result in increased P transfers to waterways (Haygarth et al., 2005; Macleod et al., 2012; Ockenden et al., 2017).

Process based models are often used to assess the response of river systems to changes in land use and future climate drivers (Bosch et al., 2014; Crossman et al., 2013; Crossman et al., 2014; El-Khoury et al., 2015; Jin et al., 2015; Whitehead et al., 2013). These models are typically considered over-parameterized, with large numbers of interacting parameters governing the key physical and biogeochemical processes represented in the model structure (Beven, 2006; Dean et al., 2009; Krueger et al., 2007). While the parameters of such models may have some physical significance, 'effective' values of those parameters are required to account for variability in the catchment, key processes and the model limitations (Beven, 1996; Beven, 2002; Beven, 2006), with these frequently estimated through a combination of manual and automated calibration procedures.

Beven (2006) also highlighted that there is often limited information in the model calibration data to effectively identify calibrated values for model parameters. For example,

infrequent water quality data collection, which does not fully pick up catchment dynamics can lead to uncertainty in P load calculations (Johnes, 2007) which then impacts on the ability of the models to simulate catchment water quality accurately (Radcliffe et al., 2009). This uncertainty, coupled with other sources of uncertainty, results in equifinality, where multiple and very different parameter sets produce an equally acceptable fit to observations (Beven, 2006). A so-called 'optimum' parameter set will not then be robust to a change in the period of calibration data. In some cases, parts of a data set may not be informative in calibrating and evaluating a model (Beven and Smith, 2015). Furthermore, the concept of equifinality has been exhibited in the observed biogeochemistry of a catchment whereby signals in the observations can be explained by a large number of interacting processes (Haygarth et al., 2012).

Understanding how well these process-based models represent the key processes in the source, mobilization and delivery continuum, will improve their ability as learning tools in helping to unravel the complex interactions occurring in a catchment. This is particularly the case where the processes are often difficult or impossible to measure at the catchment scale (e.g. phosphorus concentrations in different nutrient pools in the soil). As a result, in recent years the impact of such uncertainties has received increased attention in water quality modelling (Dean et al., 2009; Harmel et al., 2014; Karamouz et al., 2015; Page et al., 2007; Vrugt and Sadegh, 2013; Woznicki and Nejadhashemi, 2014; Yen et al., 2015).

The Generalized Likelihood Uncertainty Estimation (GLUE) methodology (Beven and Binley, 1992) is an uncertainty estimation technique widely applied in the field of environmental modelling, including water quality models (Dean et al., 2009; Krueger et al., 2010; Krueger et al., 2009; Krueger et al., 2012; Page et al., 2003; Page et al., 2007; Page et al., 2004; Rankinen et al., 2006). GLUE evaluates model realizations for acceptability in the face of uncertainty in the model structure, parameters and input data. It accepts the

equifinality concept in using a set of acceptable or behavioral models to estimate the uncertainty in model predictions. It also provides a framework to evaluate a model as fit for purpose in representing the dynamics of a catchment using a set of evaluation criteria.

In this study, GLUE is used with a 'limits of acceptability' approach to evaluate a model parameter set, which should take into account the inherent error in the calibration data, such as errors in discharge data arising from rating curve uncertainties (Blazkova and Beven, 2009; Krueger et al., 2010; McMillan et al., 2012; McMillan and Westerberg, 2015; Pappenberger et al., 2006; Westerberg et al., 2011) and errors in water quality data (Krueger et al., 2012; Page et al., 2003; Page et al., 2004; Rankinen et al., 2006). The advantage of this approach is that it allows varying limits to be set for individual observations as well as combining evaluations based on different types of observations in a consistent way (Beven, 2006). Furthermore, it has been demonstrated that high frequency coupled hydrochemical data, allows short term changes in catchment dynamics to be better captured (Benettin et al., 2015; Halliday et al., 2015) and a greater understanding of the complex and non-linear interactions in the catchment system to be obtained. This is particularly the case in flashy catchments where storm events can lead to rapid changes in stream concentrations of P, and thus allows more robust and empirically defined model evaluation criteria to be set. However, the reality of not having such high quality data available can often make it difficult to define appropriate limits (Dean et al., 2009).

The Soil and Water Assessment Tool (SWAT; Arnold et al., 1998; Gassman et al., 2007) is one such process-based model that has been the focus of uncertainty and calibration procedures in recent years (Arnold et al., 2012; Karamouz et al., 2015; Schuol and Abbaspour, 2006; Shen et al., 2012a). Designed to simulate the impacts of management and mitigation on biogeochemistry and water quality in ungauged river basins, development of SWAT began in the early 1990s (Gassman et al., 2007). The model has been continually

improved over the years and has incorporated key components based on those in other established models. These include the hydrology component from the Chemicals, Runoff, and Erosion from Agricultural Management Systems (CREAMS) model (Knisel, 1980), the pesticide component from the Groundwater Loadings Effects on Agricultural Management Systems (GLEAMS) model (Leonard et al., 1987) and the crop growth component from the Environmental Impact Policy Climate model (Izaurralde et al., 2006), which was previously known as the Erosion Productivity Impact Calculator (EPIC) model (Williams, 1990). Finally, river routing and instream kinetic routines were incorporated based around the Routing Options to Outlet (ROTO; Arnold et al., 1995) and QUAL2E (Brown and Barnwell Jr., 1987) models respectively.

The GLUE framework has been applied to SWAT before (Karamouz et al., 2015; Shen et al., 2012a) with the Nash-Sutcliffe efficiency (NSE) typically used as the likelihood measure. A prescribed threshold is used to define behavioral simulations, with focus tending to be on how the model performs in the medium to long term (typically monthly to yearly). These studies demonstrated that high uncertainty exists in the model predictions with a number of key parameters for flow and nutrient processes being unidentifiable due to limitations in the model input and calibration data (Shen et al., 2012a). However, due to limited computational power, these studies sampled only a small area of the parameter space (10,000 iterations for a 20 parameter space) and hence could miss sampling potentially behavioral parameter sets. Further to this, previous uncertainty applications to SWAT focus largely on using summary statistics such as NSE to evaluate model performance (Shen et al., 2012a; Shen et al., 2012b; Shen et al., 2013) and do not focus on those time-steps critical to model failure. Finally, whilst there have been previous studies with SWAT that are concerned with the effects of input data uncertainty on model performance (Shen et al., 2012b; Shen et al., 2013), no previous study accounts for uncertainty in the data used to calibrate the model.

This work provides for the first time, a 'limits of acceptability' approach of the GLUE framework to the SWAT model in an application to the Newby Beck sub-catchment of the River Eden Basin in Cumbria, UK. This study takes advantage of the high temporal resolution water quality monitoring data set from the Demonstration Test Catchments (DTC) project (McGonigle et al., 2014) to gain a better understanding of the uncertainty in the predictions of models such as SWAT by using the 'limits of acceptability' to identify exact time-steps critical to model failure. This will provide an insight as to whether it is suitable to use SWAT as a catchment management tool in the Newby Beck sub-catchment. We do this by evaluating whether it can adequately represent the key dynamics of P transport to the stream, whilst also explicitly accounting for errors in calibration data. This study has the following objectives.

1) What are the critical time-steps causing the model to be classed as not acceptable?
2) What can be learned from the uncertainty in the model predictions to better understand the complex interactions occurring at the catchment scale?
3) Can we identify which processes require further investigation in the model structure and do we have sufficient input data to drive such complex models?

## 2 Materials and Methods

### 2.1 Catchment description and observations

Newby Beck (Figure 1) is a small headwater sub-catchment located in the River Eden basin in the North West of England, in the United Kingdom. The catchment is approximately 12.5 $km^2$ in size with an average elevation of 234 m above sea level (Owen et al., 2012; Perks et al., 2015). The underlying geology is dominated by Carboniferous limestone, which is overlain by low-permeability glacial deposits. There are well drained, fine and loamy soils

over limestone (Waltham soil association (541q)) in the upper reaches, seasonally wet deep loamy soils in drift from Paleozoic sandstone and shale in the mid-reaches (Brickfield 3 soil association (713g) and seasonally waterlogged reddish fine and coarse loamy soils in glacial till (Clifton soil association (711n) in the lower reaches of the catchment (National Soil Resources Institute (NSRI) Cranfield University 2014). The dominant soil unit in the catchment is the 713g Brickfield association, which covers approximately 66% of the basin area. The primary land use in the catchment is improved grassland (approximately 76% by area) which is used for a mix of dairy and beef production. Other land uses are rough grassland (14%), arable (6%), woodland (2.5%) and built-over land (0.5%; Morton et al., 2011). The climate of the region is cool temperate maritime with an annual average rainfall of around 1200 mm. Due to the underlying geology, the 23% of the catchment area is greater than 5°, which results in rapid catchment response time leading to a time-to-peak of about 3 hours (Perks et al., 2015). Based on the Hydrology of Soil Types (HOST) classifications, the catchment has a standard percent runoff of 35% (Perks et al., 2015), resulting in very flashy responses of the hydrograph to rainfall events and high occurrences of saturated overland flow (Ockenden et al., 2016).

**Figure 1: Summary of spatial data in the Newby Beck catchment. Panel a) shows the catchment topography, panel b) shows the locations of the monitoring station (discharge and total phosphorus (TP)), weather station and rain gauges, panel c) shows the main soil classes in the catchment and panel d) shows the broad land use classifications.**

The catchment outlet was a rated section of channel used to provide high frequency discharge data at 15-minute intervals. The discharge measurements were calculated from a time series of stage measurements (obtained with a SWS mini-Diver) using site-specific

201 rating curves. In addition, a high frequency bankside monitoring station was situated at the
202 outlet, which recorded nitrate ($NO_3$), total P (TP) and total reactive P (TRP) at 30 minute
203 intervals (Outram et al., 2014). The TP and TRP measurements were conducted using a Hach
204 Lange combined Sigmatax sampling module and Phosphax Sigma analyzer (Perks et al.,
205 2015). Rainfall was recorded at 15-minute intervals by three tipping bucket rain gauges.
206 Other meteorological data was provided by an Automatic Weather Station (AWS), which was
207 located towards the centre of the catchment (Figure 1). Daily rainfall data was also gained
208 from a rain gauge located in the center of Newby Beck catchment from the Met Office
209 Integrated Data Archive System (MIDAS) network (Met Office 2012). The location of the
210 monitoring stations, rain gauges, and outlet monitoring station are shown in Figure 1.
211 Information on fertilizer and manure applications were based around a typical dairy and beef
212 grassland catchment system with guidance from the Defra fertilizer handbook (Rb209; Defra,
213 2013) and available farm diary data for the catchment for the years 2011-2014.
214
215 2.2 Implementation of the SWAT model to Newby Beck
216 The SWAT model (version 2012, revision 637) is a semi-distributed, process-based model
217 (Arnold et al., 1998; Gassman et al., 2007) which simulates surface and sub-surface
218 hydrology, along with various nutrient (including P) and sediment fluxes, at a basin scale.
219 The model also incorporates various land management practices along with a crop growth
220 model in order to simulate the impact of agriculture at the catchment scale. SWAT also
221 includes urban area management practices and can incorporate pollution from point sources
222 such as sewage treatment works. The model requires spatial information including land use,
223 soil type and elevation, which are often input as GIS layers. Additional inputs required
224 include any land management practices (e.g. fertilizer application rates and animal stocking
225 densities) and weather data including rainfall, temperature, wind speed, humidity and solar

226 radiation. In order to reduce the computational complexity of SWAT, a semi-distributed
227 approach is taken such that the model lumps unique land, soil and slope combinations into
228 hydrological response units (HRUs) within each sub-basin of the main catchment. The
229 hydrological and biogeochemical model processes are calculated for each HRU and then
230 lumped to produce a response for each sub-basin.
231 To implement SWAT for the Newby Beck catchment, the NextMap 5m digital elevation
232 model (DEM) dataset (Intermap Technologies 2009) was used to delineate the catchment
233 boundary highlighted in Figure 1. Land use (25 m resolution) was from the Centre of Ecology
234 and Hydrology (CEH) land cover map (LCM) 2007 (Morton et al., 2011), which indicates the
235 most likely Broad Habitat land classification for each 25m grid square. Soil properties (1 km
236 resolution) were determined from the NSRI database (Cranfield University2014). In order to
237 keep the simulation as computationally efficient as possible, the catchment was divided
238 spatially into 3 sub-basins, each with a different mean elevation. Within each sub-basin,
239 HRUs were defined based upon the unique combinations of the LCM land cover class (the
240 dominant proportion of coverage in each grid square) and the dominant soil association
241 (Brickfield (713g), resulting in 5 HRUs per sub-basin and 15 in total (Figure 1). Fertilizer
242 application rates for each land class were lumped up to HRU level to provide an average
243 nutrient application rate for each response unit. Finally, the required precipitation and
244 weather data were provided by the rain gauges and the AWS (Figure 1).
245 SWAT was set up to produce daily predictions of discharge and TP loads. A sub-daily
246 variant of the model was available (Gassman et al., 2007), however, at present it does not
247 produce sub-daily output for nutrients. Therefore in this study we have used the daily time-
248 step variant of the model which has been used in numerous previous studies (Shen et al.,
249 2012a; Shen et al., 2013; Taylor et al., 2016; Wang and Sun, 2016; Zhang et al., 2014).
250 Model simulations are evaluated using daily observations of discharge and TP loads, which

are calculated from the high frequency data at the catchment outlet. The modified SCS curve number method was used for computing surface runoff volume. While often used as a representation of infiltration excess runoff, Steenhuis et al. (1995) have shown that it can also be interpreted in terms of saturation excess contributing areas which is more appropriate for the study catchment. The Penman Monteith (Monteith, 1965) method was used to calculate evapotranspiration and the Muskingham routing method (Brakensiek, 1967; Overton, 1966) to route water in the river network. P is cycled through the soil through a combination of leaching, mineralization, decomposition and immobilization processes and surface runoff is largely assumed to be the primary transport route into the river network (Neitsch et al., 2011). The algorithms for each respective process are solved and P is moved between respective soil stores and into the river network to ensure that mass balance is conserved.

The model was run with a two year warm up period and was calibrated over the 2011-2012 and 2012-2013 hydrological years and validated over the 2013-2014 hydrological year.

2.3 The limits of acceptability GLUE uncertainty framework

The performance of the SWAT simulations was assessed using the GLUE methodology (Beven and Binley, 1992; Beven and Binley, 2014). GLUE was extended to use the limits of acceptability approach described by Beven (2006; 2009) and applied in previous applications to hydrological (Blazkova and Beven, 2009; Krueger et al., 2010; Liu et al., 2009) and water quality models (Krueger et al., 2012; Page et al., 2003; Page et al., 2004; Rankinen et al., 2006).

GLUE recognizes that for any given observational data set and performance criteria there may be multiple model parameter sets and structures that produce acceptable simulations. Each application is dependent on a number of decisions:

1. Choose which model parameters to vary

2. Choose which model structures to consider (e.g. whether in stream processing of nutrients is switched on or off)
3. Define prior distributions within which to sample each parameter
4. Determine the limits of acceptability used to assess the performance of a model run
5. Decide on a likelihood measure for creating the uncertainty prediction bounds given a set of behavioral models

In the absence of any knowledge regarding the prior probability distributions of effective parameter values, random uniform sampling was utilized between defined prior ranges. However, if this information is known it can be incorporated into the sampling strategy.

To assess if a given parameter set is behavioral, limits of acceptability are specified for each observation at each time-step during the calibration period, to take into account the inherent uncertainty in the calibration data. Model performance (*Score(t)*) is determined at each time-step, *t*, by how well the simulated value satisfies these limits and are normalized as follows to compare limits over different measures,

$$\text{Score}(t) = \begin{cases} (\hat{Y}_t - y_t)/(y_t - y_{min,t}) & \hat{Y}_t < y_t \\ (\hat{Y}_t - y_t)/(y_{max.t} - y_t) & \hat{Y}_t \geq y_t \end{cases} \quad (1)$$

where $\hat{Y}_t$ is the simulated value; $y_t$ is the best estimate of the observed value; $y_{min,t}$ is the lower limit of acceptability; and $y_{max,t}$ is the upper limit of acceptability for a given time-step. This results in scores that are zero at the best estimate of an observed value, -1 at the lower limit and +1 at the upper limit. For a model to be considered behavioral, all scores must fall within the limits at every time step (between -1 and +1).

297 The first step in defining the limits of acceptability is to consider the range of output
298 observational uncertainty. For discharges, this will depend on both water level measurement
299 uncertainty and rating curve uncertainties (e.g. McMillan and Westerberg (2015)). For water
300 quality load variables, it will depend on uncertainties in discharge, sampling and
301 measurement of determinand concentrations in addition to their aggregation to the temporal
302 and spatial scales of interest (McMillan et al., 2012). Where such uncertainties are estimated
303 using fuzzy or interval arithmetic, then limits of acceptability can be defined directly
304 (Krueger et al., 2010; Krueger et al., 2009; Krueger et al., 2012; Pappenberger et al., 2006;
305 Westerberg et al., 2011). However, where such uncertainties are estimated statistically, there
306 are normally no sharp limits on the potential ranges (the assumed distributions will have
307 infinite tails). In this case, it is necessary to truncate the uncertainty (normally at the 95% or
308 99% level).
309 Where such limits of acceptability are based only on the output observational uncertainties,
310 they provide a minimal range of acceptable behavior because no explicit account has been
311 taken of the effect of input uncertainty. This is more difficult to do since the nonlinear
312 dynamics of most models make it difficult to assess the impact of input error independently
313 of the model. There is, however, the option of exploring input error propagation within the
314 GLUE framework (Krueger et al., 2010; Krueger et al., 2009; Krueger et al., 2012; Page et
315 al., 2003; Page et al., 2004). In this paper, an indirect approach was taken by relaxing the
316 limits until a given number of behavioral simulations have been accepted. We discuss a
317 number of ways of doing so. It can be done by imposing the condition that only a certain
318 percentage of the scores must fall within the -1 to +1 scores (e.g. 95%/99%) or by finding the
319 minimum extension required of the limits for simulations to be considered behavioral. This
320 degree of relaxation can then be used to determine, at least subjectively, whether the model
321 can be considered as fit-for-purpose.

Once a set of behavioural simulations have been identified a final likelihood weight needs to be calculated for each behavioural model. First, a weight *W* is calculated at each evaluation time-step *t* using Equation 2.

$$W(t)\begin{cases}[(Score(t)-L_{lwr})/abs(L_{lwr})]^N & L_{lwr}\le Score(t)<0\\ \left[\left(L_{upr}-Score(t)\right)/abs\left(L_{upr}\right)\right]^N & 0\le Scores(t)<L_{upr}\\ 0 & Score(t)\notin\left(L_{lwr},L_{upr}\right)\end{cases} \quad (2)$$

where *Score(t)* is the normalized score at time-step *t*, and $L_{lwr}$ and $L_{upr}$ are the lower and upper criteria to consider the set of models behavioural for the required number of time steps. *N* is a shaping factor, which is set at 1 in this case, following the approach of Liu et al. (2009). This is a similar approach to applying a triangular fuzzy weight at each evaluation time-step (Freer et al., 2004; Liu et al., 2009).

The weights at each time-step are then combined to produce an overall likelihood weighting for each behavioural model:

$$L\left(M(\theta_i|Y)\right)\propto\sum_{t=1}^{T}W(t) \quad (3)$$

where *T* is the total number of time steps and *W(t)* is a triangular fuzzy weighting at time-step *t*. As previously in GLUE, prediction quantiles can then be formulated at any given time-step (*t*) by calculating the likelihood weighted cumulative density function of a predicted variable over the set of behavioural models.

$$P\left(\hat{Z}_t<z_t\right)=\sum_{j=1}^{j=N}L\left[M\left(\Theta_j\right)\middle|\hat{Z}_{t.j}<z_t\right] \quad (4)$$

where $P$ is the prediction quantile for $\hat{Z}_t$ (the simulated value of variable $Z$ at time step $t$ using model $M(\Theta_j)$) being less than $z$; $L$ is the likelihood weighting associated with model $M(\Theta_j)$; $\Theta_j$ is the $j$th parameter set; and $N$ is the number of models accepted as behavioral.

In this study, the model was evaluated using daily discharge and TP loads with the constraint imposed that for both discharge and TP loads the simulated value must fall within the limits of acceptability at all time-steps throughout the calibration period (2011-2012 and 2012-2013 hydrological years). This period totaled 731 time-steps and accounting for both upper and lower limits gave 1462 limits to satisfy for discharge. For TP loads, there were 1210 limits to satisfy, due to missing data, giving a total of 2672 limits to be met for a model run to be considered behavioural. This allows likelihood measures to be calculated for discharge ($L_Q$) and TP ($L_{TP}$), respectively. For each behavioral model run, an overall likelihood ($L_{ovr}$) can be constructed as follows

$$L_{ovr} = \frac{L_Q \cdot L_{TP}}{C} \tag{5}$$

where $C$ is a scaling factor such that the sum of likelihoods scales to unity in each case. Equation 4 can then be applied to determine the uncertainty bounds on the model predictions.

Here, thirty two parameters in the SWAT model considered important for hydrology and water quality processes (Arnold et al., 1998; Gassman et al., 2007; van Griensven et al., 2006) were sampled uniformly between the ranges detailed in the model user manual (Table 1). As some parameters varied with land use, a total of 39 were included in the Monte-Carlo simulations. In order to preserve the spatial heterogeneity of the soil and curve number parameters across HRUs, multipliers were applied during the Monte Carlo simulations (Table 1). The ranges and parameters chosen in Table 1 were based around an initial sensitivity

analysis. For such a large parameter space, many model runs were required and SWAT was implemented on the Lancaster University HEC (High End Computing) facility. The results presented are based on 5,000,000 iterations of the SWAT model executable (version 2012, revision 637), run within an R wrapper (R Core Team, 2016) which sampled the parameters uniformly between the ranges specified in Table 1.

**Table 1: SWAT model parameters and ranges used within the Generalized Likelihood Uncertainty Estimation (GLUE) framework. The values of each parameter were sampled on a random uniform basis between the ranges.**

| Parameter | Description | Min Value | Max Value |
|---|---|---|---|
| **CN2*** | SCS runoff curve number | -0.2 | 0.2 |
| **USLE_P_FRSD** | USLE[a] equation support practice factor (forest) | 0.0 | 0.5 |
| **USLE_P_AGRL** | USLE[a] equation support practice factor (arable) | 0.0 | 1.0 |
| **USLE_P_PAST** | USLE[a] equation support practice factor (pasture) | 0.0 | 0.5 |
| **USLE_P_RGRS** | USLE[a] equation support practice factor (rough grazing) | 0.0 | 1.0 |
| **USLE_P_URML** | USLE[a] equation support practice factor (urban) | 0.0 | 1.0 |
| **ALPHA_BF** | Baseflow alpha factor (1/days) | 0.0 | 1.0 |
| **GW_DELAY** | Groundwater delay (days) | 26.0 | 500.0 |
| **GWQMN** | Threshold in shallow aquifer for return flow (mm) | 970.0 | 3300.0 |
| **RCHRG_DP** | Deep aquifer percolation fraction | 0.4 | 1.0 |
| **LAT_ORGP** | Organic P in baseflow ($mgl^{-1}$) | 0.0 | 0.1 |
| **GWSOLP** | Concentration of soluble P in groundwater flow($mgl^{-1}$) | 0.0 | 0.1 |
| **GW_REVAP** | Groundwater "revap" coefficient | 0.02 | 0.2 |
| **REVAPMN** | Threshold depth in shallow aquifer for "revap" to occur (mm) | 150.0 | 500.0 |
| **SLSOIL** | Slope length for lateral subsurface flow (m) | 10.0 | 45.0 |
| **CANMX_FRSD** | Maximum canopy storage for forest ($mmH_2O$) | 0.0 | 100.0 |
| **CANMX_AGRL** | Maximum canopy storage for arable ($mmH_2O$) | 0.0 | 100.0 |
| **CANMX_PAST** | Maximum canopy storage for pasture ($mmH_2O$) | 0.0 | 100.0 |
| **CANMX_RGRS** | Maximum canopy storage for rough grazing ($mmH_2O$) | 0.0 | 100.0 |
| **LAT_TTIME** | Lateral flow travel time (days) | 0.0 | 1.8 |
| **ERORGP** | Phosphorus enrichment ratio for loading with sediment | 0.0 | 5.0 |
| **CH_N2** | Manning's "n" value for the main channel | 0.0 | 0.3 |
| **CH_COV1** | Channel erodibility factor | 0.0 | 1.0 |
| **CH_COV2** | Channel cover factor | 0.0 | 1.0 |
| **SOL_K*** | Saturated hydraulic conductivity (mm/hr) | 0.0 | 2.0 |
| **USLE_K*** | USLE[a] equation soil erodibility factor (ton $m^2$ hr/$m^3$-ton cm) | -0.1 | 0.1 |
| **SOL_ORGP** | Initial organic P concentration in soil layer ($mgl^{-1}$) | 0.1 | 100.0 |
| **SOL_LABP** | Initial labile P concentration in soil layer ($mgl^{-1}$) | 0.1 | 100.0 |
| **CH_N1** | Manning's "n" value for tributary channels | 0.06 | 0.15 |
| **SURLAG** | Surface runoff lag coefficient | 2.0 | 24.0 |
| **ESCO** | Soil evaporation compensation factor | 0.4 | 0.9 |
| **EPCO** | Plant uptake compensation factor | 0.1 | 0.9 |
| **SPEXP** | Parameter for amount of sediment reentrained in routing | 1.0 | 1.5 |

| | | | |
|---|---|---|---|
| **SPCON** | Parameter for amount of sediment reentrained in routing | 0.001 | 0.01 |
| **PSP** | P sorption coefficient | 0.01 | 0.7 |
| **CMN** | Rate factor for mineralization of organic N | 0.001 | 0.003 |
| **RSDCO** | Residue decomposition coefficient | 0.02 | 0.1 |
| **PPERCO** | P percolation coefficient (global) | 10.0 | 17.5 |
| **P_UPDIS** | P uptake distribution parameter | 10.0 | 100.0 |

*These parameters were varied relatively using a random multiplier between the ranges in order to preserve the spatial heterogeneity of the parameters.

[a]USLE= Universal Soil Loss Equation.

## 2.4 Sources of uncertainty in the calibration data

In order to set initial limits of acceptability for discharge and TP loads, the uncertainty in the rating curve and in-situ TP concentration measurements were first examined. The methodology of deriving these limits is described briefly below with more detail available in Hollaway et al (In Prep). To produce a rating curve the Velocity Area Rating Extension (VARE) model was used (Ewen et al., 2010), which uses the water balance and an assumed maximum river velocity to constrain the extrapolation of the curve beyond the gauged range. An extended version of the voting point likelihood methodology (McMillan and Westerberg, 2015) was used in a Monte Carlo Framework to calibrate the rating curve. In brief, the voting point method works by evaluating candidate rating curves (from the Monte Carlo sampling) against the observations (and in the VARE method constrained by the water balance). A candidate curve is considered behavioural if it falls within the uncertainty bounds of at least one of the observations and is weighted based upon A) the number of measurements it intersects and B) how close it lies to the true value (in this case we use a triangular weighting). Finally, 95% confidence limits are derived from all behavioural curves and their associated weightings to give the uncertainty limits on the discharge time series.

The resultant uncertainty (based on 95% prediction quantiles) on discharge was on average 96% with a range of 24-163%. This range is much larger compared to those determined during a recent study on 500 UK catchments (Coxon et al., 2015), which showed that the

majority of catchments had 20-40% relative uncertainty intervals, though the maximum uncertainty of 163% determined for Newby Beck here is much lower than the maximum value of 397% quoted by Coxon et al. (2015).

As daily TP loads are determined from both discharge and in stream TP concentrations. To evaluate the uncertainty on the in-situ concentrations, measurements from the bankside analyser were paired with land analysed grab samples and ISCO data. An empirical power law was then fitted, once again using a voting point likelihood in a Monte-Carlo framework. In this case, the lab-analysed sample was assumed representative of the true concentration. Finally, the unique combination of behavioural parameter sets from both the discharge and TP time series were used to estimate the uncertainty on the resultant TP load.

For the in-situ TP concentrations from the bankside analyser, uncertainty intervals ranged from 231% for the lower concentrations (the bottom 5%) to around 81% for the highest concentrations. When combined with the discharge uncertainty this resulted in an average 271% for the lowest loads (bottom 5%) and 76% for the highest loads.

## 3 Results

### 3.1 Model performance and rejection

For the initial limits of acceptability (see 2.4), none of the 5,000,000 parameter sets sampled produced a model that satisfied the limits at every time-step for both discharge and TP loads. In order to investigate why the sampled parameter sets were not producing behavioural models a subset of the best parameter sets was chosen on which to perform further analysis. In order to identify this subset of models we took two different approaches. These two different methods were adopted to evaluate the sensitivity of accepted model parameter sets to the choice of evaluation measure. The first approach was to find the minimum relaxation of the normalized limits across all time-steps that was required to accept a set of 1000 models. The second approach was to only require the model to fall within the

limits in the high and low flow time-steps. In this case, the thresholds for high and low flows (for both discharge and TP) were set as the top and bottom 5% of discharges as defined from the flow duration curve. For this second evaluation measure if no parameter sets satisfied the initial limits of acceptability for all the selected time steps, they were again relaxed until a set of 1000 models was accepted on which to perform further diagnostics.

3.1.1 Evaluation across all model time-steps

When the normalized scores of acceptance were allowed to relax (based on normalized scores falling within the limits at all time-steps) to $\pm$ 6.72, 1016 simulations can be considered acceptable. In order to gain a better understanding of why such large relaxation of the limits was required, a more detailed examination of the scores was made for the accepted simulations to look for systematic deviations between the simulations and observations. Figure 2 shows a summary of the performance of the 1016 simulations against observations over all time-steps, for the rising/falling limbs of the hydrograph and for the high and low flow periods (as defined above). Figure 2 also shows a comparison of the normalized scores against the observations.

**Figure 2: Generalised likelihood uncertainty estimation (GLUE) likelihood distributions, based upon the evaluation of models using criteria set for all time steps (normalized scores of ± 6.72), of Qsim (simulated discharge), normalised score for Q (discharge), TP loadsim (simulated total phosphorus) and normalised scores for TP, respectively, against observations (panels A-D). The plots are repeated for the low flow periods (panels E-H), rising time-steps (panels I-L), falling (recession) time-steps (panels M-P) and high flow periods (panels Q-T). The areas between the distribution percentiles max/min, $5^{th}$/$95^{th}$ and $25^{th}$/$75^{th}$ are shown in grey shades of increasing intensity. The medians of the distribution are shown by black dots. 1:1 lines and normalised scores of 0 lines have been added for orientation.**

For both discharge (Figure 2E) and TP loads (Figure 2G) the models tend to show a bias towards over-prediction during the low flow periods. In contrast there is systematic under-prediction shown for both discharge (Figure 2Q) and TP (Figure 2S) during the high flow

periods although the normalized scores show a tendency to be smaller for these periods which reflects the larger absolute uncertainty intervals on the higher flow observations for both measures (Figure 2). Overall, the majority of scores which tend to be outside the original limits occur during the falling limb of the time-series, particularly for the lower magnitude flows and loads during these periods, which could be a constraint on model performance.

This under-prediction of peaks during the high flow periods is reflected in Figure 3, which shows the time series of the performance of the 1016 accepted models during the summer, autumn and early winter of the 2012-2013 hydrological year. Overall, the model captures the timings of the peaks and low flow periods fairly well, however the under-prediction of the peaks in December and January is emphasized for both discharge (Figure 3a) and TP loads (Figure 3b). Despite relatively high normalized scores shown in Figure 2 during the low flow periods, the over-prediction of observations is less emphasized in Figure 3 due to the smaller absolute widths of the uncertainty intervals at these time-steps. However, over-prediction is evident during the low flow period in late January 2013, particularly in the discharge time-series.

3.1.2. Evaluation across high and low flow periods only

When the model evaluation is constrained to the high and low time-steps (top and bottom 5% of time-steps across the flow duration curve), none of the 5,000,000 model runs fall within the original limits of acceptability. Hence, in order to gain a subset of model runs for the calculation of model diagnostics, we relaxed the limits to 5.30 to gain a set of 1001 behavioural simulations. Figure 4 shows a comparison of the model performance versus the observations over all time-steps, rising/falling time-steps and high/low flow time-steps. Overall, the picture is consistent when the models were constrained over all time-steps (section 3.2.1) with over-prediction of both discharge and TP during the low flow periods (Figure 4F and 4H) and under-prediction during the high flow periods (Figure 4R and 4T).

However, much higher over-predictions are shown for lower discharge and TP loads, particularly those classified as falling time-steps (Figure 4N and 4P respectively) where normalized scores approach 15 for discharge and 30 for TP. These higher scores (compared to Figure 2) reflect the fact that we are only constraining the model on a smaller number of time-steps, albeit these are the high and low flow periods that are often considered important to simulate accurately to best capture catchment dynamics. This once again shows that poor performance during the recession periods is a constraint on finding behavioural parameter sets for SWAT in application to this catchment.

**Figure 3: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus loads (b) for Newby Beck outlet (part of the calibration period) based on normalized scores on both discharge and total phosphorus (TP) load evaluation measures when criteria (normalized scores of ± 6.72) set over all model time-steps (1016 simulations). The black line in each plot shows the observed discharge (a) and TP loads (b), respectively. The dashed lines show the uncertainty limits on the calibration data.**

**Figure 4: Generalised Likelihood Uncertainty Estimation (GLUE) likelihood distributions of, based upon the evaluation of models using criteria set for high and low flow periods only (normalized scores of ± 5.30), Qsim (simulated discharge), normalised score for Q (discharge), TP loadsim (simulated total phosphorus) and normalised scores for TP, respectively, against observations (panels A-D). The plots are repeated for the low flow periods (panels E-H), rising time-steps (panels I-L), falling (recession) time-steps (panels M-P) and high flow periods (panels Q-T). The areas between the distribution percentiles max/min, $5^{th}/95^{th}$ and $25^{th}/75^{th}$ are shown in grey shades of increasing intensity. The medians of the distribution are shown by black dots. 1:1 lines and normalised scores of 0 lines have been added for orientation.**

Figure 5 shows the time-series of model performance of the 1001 accepted models during the summer, autumn and early winter of the 2012-2013 hydrological year. In this case as the high and low flow periods that are being used to constrain the model the dynamics of the catchment are captured much better by the accepted simulations with the model capturing both the timing and magnitude of the peaks for both discharge (Figure 5a) and TP loads (Figure 5b). However, there is still under-prediction of peaks during December and early

January and over-prediction of low flow periods during late January with this once again most evident in the discharge time-series (Figure 5a).

**Figure 5: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus loads (b) for Newby Beck outlet (part of the calibration period) based on normalized scores on both discharge and total phosphorus (TP) load evaluation measures when criteria (normalized scores of ± 5.30) set over high and low flow time-steps only (1001 simulations). The black line in each plot shows the observed discharge (a) and TP loads (b), respectively. The dashed lines show the uncertainty limits on the calibration data.**

3.2 Evaluation of model parameter uncertainty

Figure 6 shows projections of the sampled points on the likelihood surface (as calculated by Equation 5) onto single parameter axes for the parameters in Table 1 for each of the behavioral simulations. These have previously been called dotty plots and can be used to infer sensitivities of the individual parameters using the Hornberger-Spear-Young method (see Beven, 2009). The points shown are the 1016 simulations which satisfy the relaxed limits of acceptability for both discharge and P when evaluated across all time-steps. The same plot is shown in Figure 7 when the models are evaluated across the high and low flow period only. Both Figures 7 and 8 show consistency in the sensitivity of the parameters varied. Of the 39 parameters varied, only four parameters exhibited any clear identifiability. These are GW_DELAY (ground water delay), RCHRG_DP (deep aquifer percolation fraction), LAT_TTIME (lateral flow travel time) and LAT_ORGP (organic P in the baseflow). Further to this, behavioural models are identified at both high and low values of the GW_DELAY parameter, which is consistent across both evaluation metrics. Some levels of identifiability were shown for the CN2 (SCS runoff curve number) and SLSOIL (slope length for lateral subsurface flow), however the responses of these parameters differed between the method chosen to evaluate the models. For SLSOIL, when the model was evaluated on all time-steps, higher likelihood values were shown towards the higher end of

the sample range. The opposite was shown for evaluation over the high and low time-steps only with higher likelihood values shown towards the lower end of the sampled parameter range. Overall the majority of parameters showed no sign of sensitivity and indicated high equifinality across the sampled ranges.

**Figure 6: Dotty plots for 39 of the parameters varied in the Monte-Carlo runs. Parameter names and definitions are shown in Table 1. These are based on the 1016 behavioral simulations evaluated across all time-steps (normalized scores of ± 6.72).**

**Figure 7: Dotty plots for 39 of the parameters varied in the Monte-Carlo runs. Parameter names and definitions are shown in Table 1. These are based on the 1001 behavioral simulations evaluated across the high and low flow time-steps only (normalized scores of ± 5.30).**

The parameters that exhibit sensitivity are all linked to runoff and sub-surface processes and all interact to affect the time taken for water to reach the river network, and thus affect the transport of P. However, the high equifinality in the other parameters (particularly those in relation to the levels of P in the soils SOL_ORGP and SOL_LABP) indicates that given the present assumptions and data available for the catchment, there is not enough information to calibrate these parameters effectively.

3.3 Critical time-steps for model failure

Figure 8 shows a breakdown of the classification (high/low or rising/falling) of the time-steps of the sub-sample of models chosen on which to perform model diagnostics that result in model failure (lie outside the original limits of acceptability). For both evaluation measures used in this study, the falling limb time-steps contribute the largest proportion of failing time-steps for both simulated discharge (37% for all time-steps evaluation and 34% for evaluation on high/low time-steps) and TP loads (30% and 50% respectively). All other time step classifications contribute roughly the same to model failure with the rising limb and high

564 flow time-steps accounting for approximately 10-15% of failures for both discharge and TP.
565 For discharge, the low flow time-steps account for around 10% of failures. However, for TP
566 loads they provide a much smaller contribution at around 3-4% indicating that model
567 performance at these time-steps may be less of a constraint on model performance for TP.
568 Overall it is shown that despite using two different model evaluation measures to accept
569 behavioural models, the falling limb time-steps are consistently shown to be a constraint on
570 model performance in this SWAT application to Newby Beck.

571 3.4 Model validation.

572 The 1016/1001 behavioral simulations (all time steps evaluation/high and low flows
573 evaluation) were then used to predict the discharge and P loads for a period not used in
574 calibration (winter of the 2013-2014 hydrological year due to data availability) in order to
575 validate the model performance (Figures 9 and 10). For discharge (Figures 9a and 10a), the
576 picture was somewhat similar during the validation period where the model tended to pick
577 out the timings of the peaks and recession periods well. Overall, under-prediction of the
578 observed discharge peaks was seen throughout the validation period being most evident
579 during mid-December 2013 and early January 2014. As when calibrating the model, the
580 under prediction of peaks was more pronounced when the models were evaluated across all
581 time-steps (Figure 9a). Both the timing and magnitude of the peaks was picked up much
582 better when constraining the models on the high and low flow periods (Figure 10a). As in
583 calibrating the model, the low flow periods were typically over-predicted by the model (on
584 both evaluation measures) with this being most evident towards the end of January 2014.

585 **Figure 8: Breakdown of classification of time-steps resulting in model failure for the**
586 **1016 simulations constrained on all time-steps (upper panel) and the 1001 simulations**
587 **constrained on the high and low flow periods only (lower panel). The bars show the**
588 **median % contribution to failing time-steps and the error bars show the 2.5/97.5$^{th}$**
589 **percentiles from the Generalised likelihood uncertainty estimation (GLUE) weighted**
590 **distributions.**
591

For TP loads, the picture is the same as during calibration with the model under-predicting all peaks, particularly when they were constrained using all time-steps where the model failed to capture the magnitude of any peak (Figures 9b and 10b). When constrained on the high and low flows time-steps only, the model reproduced the magnitudes and timings of the majority of the peak loads, however there are still cases where the model under predicts a peak by up to 75% (15$^{th}$ December 2013). Further to this the uncertainty bounds on the model predictions are much wider during the recession limbs of the TP time series, and shows over-prediction of the observations during this period.

**Figure 9: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus (TP) loads (b) for Newby Beck outlet during the validation period (winter of the 2013-2014 Hydrological year) using the 1016 behavioral simulations accepted on both discharge and total phosphorus load criteria when evaluating constrained across all time-steps. The black line in each plot shows the observed discharge (a) and TP loads (b), respectively. The dashed lines show the uncertainty limits on the calibration data.**

## 4 Discussion

This work, presents for the first time, a 'limits of acceptability' GLUE uncertainty analysis of the widely used SWAT model, using continuous high frequency water quality measurements. It was shown that when initial limits of acceptability (based upon the uncertainty in the outlet data for the calibration period), are accounted for and given the assumptions detailed, none of the 5,000,000 simulations provided suitable predictability of the dynamics of the catchment (i.e. none of them were classed as behavioral).

**Figure 10: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus (TP) loads (b) for Newby Beck outlet during the validation period (winter of the 2013-2014 Hydrological year) using the 1001 behavioral simulations accepted on both discharge and total phosphorus load criteria when evaluating constrained across high and low flow time-steps only. The black line in each plot shows the observed discharge (a) and TP loads (b), respectively.**

624 Therefore, in order to obtain behavioral simulations to investigate the uncertainty in the
625 SWAT model predictions, a subset of samples was obtained on which to perform further
626 diagnostics, with this subset chosen using two different criteria. The first was to find the
627 minimum level of relaxation across all model time-steps in the calibration period required to
628 consider the models acceptable. In this case relaxation of the limits to ±6.72 gave a subset of
629 1016 acceptable models. In the second case, we only required the models to fall within the
630 relaxed limits during periods of high and low flow (here defined as the top and bottom 5% of
631 discharges based on the flow duration curve). For these criteria, the limits had to be relaxed
632 (over the high and low flow periods only) ±5.30 to give a subset of 1001 accepted models.
633 This was across both discharge and TP loads.
634 Using these two different evaluation measures produced two distinctly different time
635 series when the models were compared with observations (Figures 5 and 7) and during the
636 validation period (Figures 9 and 10). When the models were constrained to fit within the
637 limits across all time-steps the parameter sets that are considered acceptable consistently
638 under predict the peaks in both discharge and TP loads, particularly during the validation
639 period. In contrast, when we only constrain the model on the low and high flow periods, the
640 simulations from the accepted parameter sets produce a much better representation of the
641 catchment dynamics, particularly in the magnitudes of the TP load peaks. However,
642 constraining the model in this way accepts simulations that have poor performance during the
643 rising limb and recession periods where the normalized scores approach 15 in the case of
644 discharge and 30 in the case of TP loads. This contrast between the chosen metric to evaluate
645 the model is the result of several different factors and depends on the characteristics and
646 dynamics of the Newby Beck catchment. Due to its flashy nature and low baseflow index
647 (Ockenden et al., 2016; Outram et al., 2014), Newby Beck is dominated by sub-daily
648 processes which may lead to timing errors in the simulated hydrograph from SWAT due to

649 the use of the daily time-step of the model. Therefore, when all time-steps are included in the
650 evaluation metric, there is a high chance of the model simulations producing high normalized
651 scores. However, as reported recently by (Coxon et al., 2014), constraining the model using
652 time-step measures such as these can be a very critical test of the model, particularly due to
653 the strong influence of observational uncertainty on such metrics (see Section 3.1). This is
654 shown in Figure 3 where all of the accepted 1016 simulations (when using the all-time-step
655 metric) under-predicted the peaks by a large amount for both discharge and TP loads, despite
656 being considered acceptable within the relaxed limits of 6.72. This could be because the
657 normalized scores are based upon the relative uncertainty intervals around the observations,
658 which allows a larger absolute deviation from the observed value on the peaks. This is a case
659 of accepting a model that is not a good representation of the processes but which fits within
660 the errors in the calibration data (Beven, 2012; Beven and Smith, 2015). It should also be
661 noted that the normalized scores are also based on estimates of the 95% limits around each
662 observation (see 2.4) and therefore the potential range of uncertainty could be larger. In order
663 to test the effect of this on model evaluation, we performed the same analysis of relaxing the
664 scores until 1057 simulations were accepted. However, in this instance we only required the
665 model to fit the limits at 95% of the time-steps. Figure 11 shows the time series of discharge
666 and TP compared to the observations and shows that when accounting for the model only
667 fitting the time-steps 95% of the time, the model still produces simulations where the peaks
668 are underestimated, such as in early January 2013. Hence, there is the still the risk of poor
669 models being accepted due to uncertainty in the calibration data.

670 **Figure 11: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction**
671 **bounds (green shading) for discharge (a) and total phosphorus loads (b) for Newby**
672 **Beck outlet (part of the calibration period) based on normalized scores on both**
673 **discharge and total phosphorus (TP) load evaluation measures when criteria set over**
674 **95% of time steps (1057 simulations). The black line in each plot shows the observed**
675 **discharge (a) and TP loads (b), respectively.**
676

When the lesser constraint of just high and low flows (often the periods of most nutrient transport in flashy catchments (Haygarth et al., 2005; Ockenden et al., 2016; Perks et al., 2015)) was applied simulations that match the peaks and low flow periods with a greater degree of accuracy were produced. This also required less relaxation of the limits of acceptability (± 5.30). This is in agreement with the recent work of (Coxon et al., 2014) showing that the performance of behavioural models accepted using different diagnostics can be strongly linked to the dominant processes occurring in the catchment. In this case, we have shown that constraining the models on high and low flow periods only in a flashy catchment produces a model ensemble that captures the peak discharges and TP loads better. However, the utilization of this diagnostic further highlights the time-steps resulting in poor model performance, where time-steps not used in the evaluation (e.g. the rising and falling time-steps) return much higher normalized scores (in excess of 30 as shown in Figure 5) than when the metric across all time-steps is used.

However, we have shown here that, despite the choice of evaluation metric, a consistent picture emerges about which class of time-step is contributing most to model failures (Figure 8). Overall, the falling limb/recession time-steps were consistently a constraint on model performance contributing between 30-50% of failing time-steps for discharge and TP time-steps across both evaluation measures. This therefore indicates potential errors in the model structure of SWAT of the representation of sub-surface processes, an area of the model that has been shown to perform poorly in the past (Guse et al., 2014).

For a large number of the parameters, it is difficult to identify any sensitivity in fitting the observations, and a large amount of equifinality is evident (Figures 7 and 8). This is particularly the case for the SOL_ORGP (soil organic P) and SOL_LABP (soil labile P) parameters, which show no clear sensitivity at all using the likelihood measure based on the limits of acceptability. Both of these parameters have been shown to play an important role in

the amount of P in the water course and are often very difficult to measure in any detail at the catchment scale (Schoumans et al., 2009). It is accepted that given a 39 dimension parameter space, 5,000,000 SWAT runs provides only a small sample of the model parameter space, albeit many more than any previously published SWAT calibration exercise, and that such a small sample can contribute to the uncertainty. Thus, there is the possibility of missing potentially behavioral models during the sampling process. They are clearly, however, sparsely distributed even with the relaxed limits of acceptability. Further adding to model parameter uncertainty is the GW_DELAY parameter, which exhibits strong identifiability, but showing the identification behavioural models at both extremes of the parameter range. Therefore in this application of SWAT both high and low groundwater delay times produce equivalent model performance in terms of the relaxed limits of acceptability. This infers that there could be compensation processes occurring in the sub-surface module of the model or could highlight additional issues in the model structural representation of groundwater attenuation in the catchment.

The limits of acceptability approach provides advantages over more traditional evaluation metrics such as NSE and root mean square error (RMSE). These are global measures, which tend to focus on the average error from the data over the calibration period, rather than focus on the individual time-steps that are causing the model to fail. The limits approach utilizes the high frequency data to provide a more detailed evaluation of the model and allows the identification of critical time-steps that are causing poor model performance. Further to this, the limits approach goes someway to accounting for uncertainty in the data/observations used to calibrate the model.

However, it is impossible to make this method completely objective due to the difficulty in accounting for error in the model inputs. In past applications of the GLUE limits of acceptability approach (Liu et al., 2009) the relaxation of the limits was justified to account

for uncertainty in the model input data. However, in this case the model user must examine the degree of relaxation in the scores and utilize the available knowledge of the inputs to see if the level of relaxation is acceptable. Given the epistemic nature of the input uncertainties, it is difficult to truly assess the effect of input error and its representation needs to be independent of the model structure (e.g. Beven, 2006). One method is to employ the use of an statistical error model to account for input error in the model (e.g. Krueger et al. (2010), go some way to accounting for this) but it is difficult to create a realistic error model, even for rainfall inputs. It would also be even more computationally expensive and thus was not implemented in the present work.

The effects of both input error and model structural errors should be seen in the deviations outside the normalised limits. The results show that the limits have to be relaxed by a very large amount (up to a factor of 6.72) to gain a set of behavioral simulations that allows the sensitivity of the parameter sets to be explored. An examination of the potential input errors to the catchment system has been taken in this study to determine whether a relaxation by factors of up to seven are acceptable. In the Newby Beck catchment, there are four rain gauges sited in a relatively small area (12.5 $km^2$ – Figure 1). It is still possible that some rainfall in the catchment could be missed in the model input, particularly during summer convective storms, leading to commensurability issues with the rainfall input (Beven and Smith, 2015; Beven et al., 2011). Different rainfall input realizations and associated errors have previously been shown to impact model performance (Blazkova and Beven, 2009). However, due to the relatively good coverage by the rain gauges in the Newby Beck catchment, errors in the rainfall input are likely to be small. It can therefore be concluded that it is model structural error, rather than input error, that is leading to the high relaxation of the limits required to define model realisations of the hydrograph as acceptable.

751 With respect to P, there is a much larger uncertainty in the overall inputs into the catchment,
752 particularly to the exact amounts of fertilizer spread on the land and the amount of dung
753 deposited from grazing. Lacking more detailed information, the inputs used in this
754 application of SWAT are based upon Defra recommendations (Defra, 2013) and local
755 knowledge of the catchment. Furthermore, the lumped nature of the SWAT model requires
756 average P inputs for each HRU, which can add further uncertainty in the amount of nutrients
757 added to the system. This can therefore lead to the locations of the inputs being smoothed out
758 leading to commensurability issues. However, the average amount of P added to the
759 catchment per year during the run (2.3 kg $ha^{-1}$) is much smaller than the levels of P in the soil
760 stores during the course of the run (approximately 15000 kg $ha^{-1}$). Thus, errors in P inputs
761 and timing are unlikely to have an effect on the levels of P being transported to the stream
762 compared to uncertainty and errors in the parameters and model structures, which govern the
763 mobilisation and transport of P in the soil. Previous work on similar small-sized catchments
764 also suggests that hydrological and biochemical processes have a much larger control on the
765 temporal variations in stream P in the catchment, rather than the timings and magnitudes of
766 the agricultural inputs (Dupas et al., 2015; Haygarth et al., 2012). In this work, we explicitly
767 account for the uncertainty in soil P by varying the SOL_ORGP and SOL_LABP (organic
768 and labile P soil stores) as part of the GLUE analysis with both of these parameters showing
769 high equifinality. It has also been shown in previous analysis on Newby Beck (Ockenden et
770 al., 2016), that the observed TP loads during storm events in the catchment are highly
771 correlated with peaks in rainfall. These storm events account for approximately 83% of the
772 annual TP load indicating that rainfall plays a strong role in controlling the transport of TP
773 into the stream network. As discussed above, the errors in rainfall are likely to be relatively
774 low in this catchment, and given its importance as a driver of TP transport along with the
775 small contribution of P inputs to overall soil P, we can conclude that relaxing the limits by a

776 factor of 6.72 is not acceptable in this application of SWAT to Newby Beck. We can
777 therefore conclude that, as with discharge, model structural error is the likely cause of this
778 requirement to relax the constraints by such a substantial amount.
779 The ability of the model to adequately simulate the observed TP loads is also further
780 compounded by the poor performance of SWAT in terms of discharge evaluation, given that
781 discharge is part of the TP load calculation. Hence, as model structural error has been shown
782 to be such a large constraint in the accurate prediction of discharge and thus TP loads, it is
783 unlikely that improvements in input data will greatly improve model predictions. In addition
784 to this, even in a small experimental catchment, gaining sufficient improvement in model
785 input data would require significant expense. In the case of TP, this would require detailed
786 farmer logs in timings and location of fertilizer applications, detailed monitoring of surface
787 and subsurface storage and availability of TP in the catchment, along with detailed field scale
788 budgets of the nutrients in the soils.
789 This prompts an additional question, if we are required to relax the limits, which are
790 primarily due to structural error in the model, by a factor of 6.72, should we go to the expense
791 of collecting the additional input data required by such a complex model structure? It has
792 been shown in previous work (Dean et al., 2009; Shen et al., 2012a) that insufficient input
793 data are a constraint on even the best of models, therefore clearly improvement is required on
794 both sides. The advantage of using the limits of acceptability approach is that we can use the
795 results of the model evaluation to target which areas of the model structure require
796 improvement and infer which areas are best to target our efforts for additional data collection,
797 particularly in situations where funds for such efforts are limited.
798
799 **5 Conclusions**

800 This study has presented the first 'limits of acceptability' assessment of the SWAT model
801 using continuous high frequency discharge and water quality monitoring data. We highlight
802 that having the availability of high frequency data coupled with the GLUE 'limits of
803 acceptability' approach; the model performance can be assessed taking into account the
804 uncertainty on the calibration data at each time-step. This provides greater insights into why
805 the model is failing beyond the more traditional global measures of model evaluation such as
806 NSE and RMSE.
807 In the application of SWAT to the Newby Beck headwater catchment in the UK, it is
808 shown that the limits of acceptability based on output observational uncertainties have to be
809 relaxed by a substantial amount (by factors of between 5.3 and 6.72 on a normalized scale
810 depending on the evaluation criteria used) in order to produce a set of behavioral simulations
811 (1001 and 1016 respectively out of 5,000,000 realizations) on which to perform model
812 diagnostics. In this case, despite the evaluation metric used, the model is shown to
813 consistently perform poorly during periods of recession in both the discharge and TP time
814 series, with uncertainty in the representation of subsurface flow pathways identified as a
815 potential cause for this poor performance. During the validation period the model was shown
816 to capture the timings of peaks in the river TP load, however, it was shown to often predict
817 the magnitude of these peaks poorly. This work raises an interesting point- how much
818 relaxation is allowable in the limits of acceptability before we consider the model as not
819 providing useful predictions of the processes occurring in the catchment? On the one hand,
820 we have learnt from the model to identify areas where we need to focus future model
821 development and data collection efforts in river catchments. On the other, we have shown
822 that in this particular case, SWAT is not fit for purpose to be used as a management tool due
823 to the large uncertainty bounds on predictions, particularly during the validation period. This
824 conclusion agrees with previous applications of SWAT to other catchments of similar

catchment areas and similar geoclimatic circumstances (Hoang et al., 2017; Moges et al., 2017; Schneiderman et al., 2007). Therefore, despite being used in numerous catchments worldwide (often with less rigorous evaluation), SWAT may not be fit for purpose as a general management tool, particularly in flashy catchments being dominated by overland flow where the model structure may be inadequate to accurately capture the major catchment processes dominating P transfer.

However, there is still a need to advise policy makers on how changes in the environment are likely to affect hydrology and water quality in the future and what mitigation measures to take, if any. A number of potential options are available, such as precautionary methods suggested by Beven (2011), or the use of fuzzy modelling methods (Page et al., 2012; Zhang et al., 2013) or finding another process based model to use – though it is highly likely that another model will suffer the same uncertainty issues as shown here with SWAT. A final option is to shift towards more simple P transfer model (E.g. Dupas et al. (2016)) which have been shown to capture P losses well with minimum input data. However as highlighted by Dupas et al. (2016), such models still have uncertainties associated with them and in some cases still require substantial relaxation of the 'limits of acceptability'.

We acknowledge that process-based models may be potentially useful catchment management tools. They are often used to quantify the effects of changes in catchment conditions (e.g. climate change) on the behavior of nutrients in catchments (Crossman et al., 2014; Wang and Sun, 2016). They are primarily used because they provide a numerical representation of conceptual processes that in theory represent how these processes adapt to changing environmental conditions under different scenarios. However, the results presented here stress the importance of having the best available input data along with high frequency data from continuous monitoring systems for rigorous model evaluation, as highlighted in previous studies (Benettin et al., 2015; Dupas et al., 2016; Halliday et al., 2015; Ockenden et

al., 2017). High frequency data allows us to set more robust 'limits of acceptability', particularly in catchments with a flashy response where infrequent grab samples may fail to capture key processes/events and may not provide a stringent enough test of the model structure/processes. The results also imply that more needs to be done to improve the ability of the model to simulate the dynamics of key catchment processes with parameters that are more identifiable in practical applications, or more easily estimated in predicting future conditions. Finally, our results also indicate the possibility that even with the best representation of the key processes in the model structure; we still may have a long way to go to have sufficient input data to adequately drive such complex model structures.

The study has not resolved the issue of how far the limits of acceptability should be relaxed to provide a set of models considered useful for predicting outcomes. That is a question for individual users to consider for particular types of applications, i.e. can we be objective about the effects of input error on model performance, particularly for predicting nutrient responses? This study suggests that SWAT may not be fit-for-purpose in this particular application, however, confirmation of its general applicability, or not, requires critical testing of the method on multiple models and multiple catchment datasets in ways that allow for uncertainty and potential equifinality of model representations.

**Acknowledgments**

This study was funded by the Natural Environment Research Council (NERC) as part of the NUTCAT 2050 project, grants NE/K002392/1, NE/K002430/1 and NE/K002406/1, and supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101). The authors are grateful to the Eden Demonstration Test Catchment (Eden DTC) research platform for provision of the field data (Department for Environment, Food and Rural Affairs (Defra), projects WQ0210, WQ0211, WQ0212 and LM0304). The data

used in this study are openly available from the Lancaster University data archive (details reserved until publication). The DTC data are available from the Eden DTC consortium until the data archive is transferred to Defra (Department for Environment, Food & Rural Affairs) as the holding body. The SWAT model executable and source code are open source and are available for download at http://swat.tamu.edu/.

## References


Arnold, J.G. et al., 2012. Swat: Model Use, Calibration, and Validation. Transactions of the Asabe, 55(4): 1491-1508.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment - Part 1: Model development. Journal of the American Water Resources Association, 34(1): 73-89. DOI:10.1111/j.1752-1688.1998.tb05961.x

Arnold, J.G., Williams, J.R., Maidment, D.R., 1995. Continuous-time water and sediment routing model for large basins. Journal of Hydraulic Engineering-Asce, 121(2): 171-183. DOI:10.1061/(asce)0733-9429(1995)121:2(171)

Benettin, P., Kirchner, J.W., Rinaldo, A., Botter, G., 2015. Modeling chloride transport using travel time distributions at Plynlimon, Wales. Water Resour. Res., 51(5): 3259-3276. DOI:10.1002/2014WR016600

Beven, K., 1996. A discussion of distributed hydrological modelling. In: JC, M.B.R.A. (Ed.), Distributed hydrological modelling. Kluwer, Netherlands, pp. 255-278.

Beven, K., 2002. Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. Hydrological Processes, 16(2): 189-206. DOI:10.1002/hyp.343

898 Beven, K., 2006. A manifesto for the equifinality thesis. J. Hydrol., 320(1-2): 18-36.
899 DOI:10.1016/j.jhydrol.2005.07.007

900 Beven, K., 2009. Environmental modelling: an uncertain future? Routledge, London.

901 Beven, K., 2011. I believe in climate change but how precautionary do we need to be in
902 planning for the future? Hydrological Processes, 25(9): 1517-1520.
903 DOI:10.1002/hyp.7939

904 Beven, K., 2012. Rainfall-runoff modelling: the primer. Wiley-Blackwell, Chichester.

905 Beven, K., Binley, A., 1992. The future of distributed models - model calibration and
906 uncertainty prediction. Hydrological Processes, 6(3): 279-298.
907 DOI:10.1002/hyp.3360060305

908 Beven, K., Binley, A., 2014. GLUE: 20 years on. Hydrological Processes, 28(24): 5897-5918.
909 DOI:10.1002/hyp.10082

910 Beven, K., Smith, P., 2015. Concepts of Information Content and Likelihood in Parameter
911 Calibration for Hydrological Simulation Models. J. Hydrol. Eng., 20(1): 15.
912 DOI:10.1061/(asce)he.1943-5584.0000991

913 Beven, K., Smith, P.J., Wood, A., 2011. On the colour and spin of epistemic error (and what
914 we might do about it). Hydrology and Earth System Sciences, 15(10): 3123-3133.
915 DOI:10.5194/hess-15-3123-2011

916 Blazkova, S., Beven, K., 2009. A limits of acceptability approach to model evaluation and
917 uncertainty estimation in flood frequency estimation by continuous simulation: Skalka
918 catchment, Czech Republic. Water Resour. Res., 45: 12. DOI:10.1029/2007wr006726

Bosch, N.S., Evans, M.A., Scavia, D., Allan, J.D., 2014. Interacting effects of climate change and agricultural BMPs on nutrient runoff entering Lake Erie. J. Gt. Lakes Res., 40(3): 581-589. DOI:10.1016/j.jglr.2014.04.011

Brakensiek, D.L., 1967. Kinematic Flood Routing. Transactions of the ASAE, 10(3): 340-343.

Brown, L.C., Barnwell Jr., T.O., 1987. The enhanced water quality models QUAL2E and QUAL2E-UNCAS: Documentation and user manual. EPA document EPA/600/3-87/007., USEPA, Athens GA.

Coxon, G., Freer, J., Wagener, T., Odoni, N.A., Clark, M., 2014. Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. Hydrological Processes, 28(25): 6135-6150. DOI:10.1002/hyp.10096

Coxon, G. et al., 2015. A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. Water Resour. Res., 51(7): 5531-5546. DOI:10.1002/2014wr016532

Cranfield University, 2014. The Soils Guide, Cranfield University.

Crossman, J. et al., 2013. Impacts of climate change on hydrology and water quality: Future proofing management strategies in the Lake Simcoe watershed, Canada. J. Gt. Lakes Res., 39(1): 19-32. DOI:10.1016/j.jglr.2012.11.003

Crossman, J. et al., 2014. Flow pathways and nutrient transport mechanisms drive hydrochemical sensitivity to climate change across catchments with different geology and topography. Hydrol. Earth Syst. Sci., 18(12): 5125-5148. DOI:10.5194/hess-18-5125-2014

941 Dean, S., Freer, J., Beven, K., Wade, A.J., Butterfield, D., 2009. Uncertainty assessment of a
942 process-based integrated catchment model of phosphorus. Stoch. Environ. Res. Risk
943 Assess., 23(7): 991-1010. DOI:10.1007/s00477-008-0273-z

944 Defra, 2013. The British Fertiliser survey of fertiliser practice. Fertiliser use on farm crops for
945 crop year 2012.

946 Dupas, R., Gascuel-Odoux, C., Gilliet, N., Grimaldi, C., Gruau, G., 2015. Distinct export
947 dynamics for dissolved and particulate phosphorus reveal independent transport
948 mechanisms in an arable headwater catchment. Hydrological Processes, 29(14): 3162-
949 3178. DOI:10.1002/hyp.10432

950 Dupas, R. et al., 2016. Uncertainty assessment of a dominant-process catchment model of
951 dissolved phosphorus transfer. Hydrol. Earth Syst. Sci., 20(12): 4819-4835.
952 DOI:10.5194/hess-20-4819-2016

953 El-Khoury, A. et al., 2015. Combined impacts of future climate and land use changes on
954 discharge, nitrogen and phosphorus loads for a Canadian river basin. Journal of
955 Environmental Management, 151: 76-86. DOI:10.1016/j.jenvman.2014.12.012

956 European Union, 2000. Directive 2000/60/EC: Establishing a framework for Community
957 action in the field of water policy (The Water Framework Directive).

958 Ewen, J., Geris, J., O'Donnell, G., Mayes, Q., O'Connell, E., 2010. Multiscale Experimentation,
959 Moitoring and Analysis of Long-term Land Use Changes and Flood Risk - SC060092:
960 Final Science Report, Newcastle University, Newcastle-Upon-Tyne.

961 Freer, J.E., McMillan, H., McDonnell, J.J., Beven, K.J., 2004. Constraining dynamic
962 TOPMODEL responses for imprecise water table information using fuzzy rule based

963 performance measures. J. Hydrol., 291(3–4): 254-277.
964 DOI:https://doi.org/10.1016/j.jhydrol.2003.12.037

965 Gassman, P.W., Reyes, M.R., Green, C.H., Arnold, J.G., 2007. The soil and water assessment
966 tool: Historical development, applications, and future research directions. Transactions
967 of the Asabe, 50(4): 1211-1250.

968 Guse, B., Reusser, D.E., Fohrer, N., 2014. How to improve the representation of hydrological
969 processes in SWAT for a lowland catchment – temporal analysis of parameter
970 sensitivity and model performance. Hydrological Processes, 28(4): 2651-2670.
971 DOI:10.1002/hyp.9777

972 Halliday, S.J. et al., 2015. High-frequency water quality monitoring in an urban catchment:
973 hydrochemical dynamics, primary production and implications for the Water
974 Framework Directive. Hydrological Processes, 29(15): 3388-3407.
975 DOI:10.1002/hyp.10453

976 Harmel, R.D. et al., 2014. Evaluating, interpreting, and communicating performance of
977 hydrologic/water quality models considering intended use: A review and
978 recommendations. Environ. Modell. Softw., 57: 40-51.
979 DOI:10.1016/j.envsoft.2014.02.013

980 Haygarth, P.M. et al., 2012. Scaling up the phosphorus signal from soil hillslopes to headwater
981 catchments. Freshwater Biology, 57: 7-25. DOI:10.1111/j.1365-2427.2012.02748.x

982 Haygarth, P.M., Wood, F.L., Heathwaite, A.L., Butler, P.J., 2005. Phosphorus dynamics
983 observed through increasing scales in a nested headwater-to-river channel study. Sci.
984 Total Environ., 344(1-3): 83-106. DOI:10.1016/j.scitotoenv.2005.02.007

985 Hoang, L. et al., 2017. Predicting saturation-excess runoff distribution with a lumped hillslope
986 model: SWAT-HS. Hydrological Processes, 31(12): 2226-2243.
987 DOI:10.1002/hyp.11179

988 Intermap Technologies, 2009. NEXTMap British Digital Terrain (DTM) Model Data by
989 Intermap. NERC Earth Observation Data Centre.

990 Izaurralde, R.C., Williams, J.R., McGill, W.B., Rosenberg, N.J., Jakas, M.C.Q., 2006.
991 Simulating soil C dynamics with EPIC: Model description and testing against long-
992 term data. Ecological Modelling, 192(3-4): 362-384.
993 DOI:10.1016/j.ecolmodel.2005.07.010

994 Jin, L. et al., 2015. Assessing the impacts of climate change and socio-economic changes on
995 flow and phosphorus flux in the Ganga river system. Environ. Sci.-Process Impacts,
996 17(6): 1098-1110. DOI:10.1039/c5em00092k

997 Johnes, P.J., 2007. Uncertainties in annual riverine phosphorus load estimation: Impact of load
998 estimation methodology, sampling frequency, baseflow index and catchment
999 population density. J. Hydrol., 332(1-2): 241-258. DOI:10.1016/j.jhydrol.2006.07.006

1000 Jones, P., Harpham, C., Kilsby, G.C., Glenis, V., Burton, A., 2010. UK Climate Projections
1001 science report: Projections of future daily climate for the UK from the Weather
1002 Generator, Met Office, UK.

1003 Karamouz, M., Taheriyoun, M., Seyedabadi, M., Nazif, S., 2015. Uncertainty based analysis
1004 of the impact of watershed phosphorus load on reservoir phosphorus concentration. J.
1005 Hydrol., 521: 533-542. DOI:10.1016/j.jhydrol.2014.12.028

1006 Kendon, E.J. et al., 2014. Heavier summer downpours with climate change revealed by weather
1007 forecast resolution model. Nat. Clim. Chang., 4(7): 570-576.
1008 DOI:10.1038/nclimate2258

1009 Knisel, W.G., 1980. CREAMS: A field scale model for chemical, runoff, and erosion from
1010 agricultural management system, Conservation Research Report No. 26, U.S.
1011 Department of Agriculture, Washington DC.

1012 Krueger, T., Freer, J., Quinton, J.N., Macleod, C.J.A., 2007. Processes affecting transfer of
1013 sediment and colloids, with associated phosphorus, from intensively farmed grasslands:
1014 a critical not on modelling phosphorus transfers. Hydrological Processes, 21(4): 557-
1015 562. DOI:10.1002/hyp.6596

1016 Krueger, T. et al., 2010. Ensemble evaluation of hydrological model hypotheses. Water Resour.
1017 Res., 46(7): n/a-n/a. DOI:10.1029/2009WR007845

1018 Krueger, T. et al., 2009. Uncertainties in Data and Models to Describe Event Dynamics of
1019 Agricultural Sediment and Phosphorus Transfer All rights reserved. No part of this
1020 periodical may be reproduced or transmitted in any form or by any means, electronic
1021 or mechanical, including photocopying, recording, or any information storage and
1022 retrieval system, without permission in writing from the publisher. J. Environ. Qual.,
1023 38(3): 1137-1148. DOI:10.2134/jeq2008.0179

1024 Krueger, T. et al., 2012. Comparing empirical models for sediment and phosphorus transfer
1025 from soils to water at field and catchment scale under data uncertainty. European
1026 Journal of Soil Science, 63(2): 211-223. DOI:10.1111/j.1365-2389.2011.01419.x

1027 Leonard, R.A., Knisel, W.G., Still, D.A., 1987. GLEAMS - Groundwater Loading Effects of
1028 Agricultural Management-Systems. Transactions of the Asae, 30(5): 1403-1418.

1029 Liu, Y.L., Freer, J., Beven, K., Matgen, P., 2009. Towards a limits of acceptability approach to
1030 the calibration of hydrological models: Extending observation error. J. Hydrol., 367(1-
1031 2): 93-103. DOI:10.1016/j.jhydrol.2009.01.016

1032 Macleod, C.J.A., Falloon, P.D., Evans, R., Haygarth, P.M., 2012. The Effects of Climate
1033 Change on the Mobilization of Diffuse Substances from Agricultural Systems. In:
1034 Sparks, D.L. (Ed.), Advances in Agronomy, Vol 115. Advances in Agronomy, pp. 41-
1035 77. DOI:10.1016/b978-0-12-394276-0.00002-0

1036 McGonigle, D.F. et al., 2014. Developing Demonstration Test Catchments as a platform for
1037 transdisciplinary land management research in England and Wales. Environ. Sci.-
1038 Process Impacts, 16(7): 1618-1628. DOI:10.1039/c3em00658a

1039 McMillan, H., Krueger, T., Freer, J., 2012. Benchmarking observational uncertainties for
1040 hydrology: rainfall, river discharge and water quality. Hydrological Processes, 26(26):
1041 4078-4111. DOI:10.1002/hyp.9384

1042 McMillan, H.K., Westerberg, I.K., 2015. Rating curve estimation under epistemic uncertainty.
1043 Hydrological Processes, 29(7): 1873-1882. DOI:10.1002/hyp.10419

1044 Met Office, 2012. Met Office Integrated Data Archive System (MIDAS) Land and Marine
1045 Surface Stations Data (1853-current). . In: NCAS British Atmospheric Data Centre
1046 (Ed.).

1047 Moges, M.A. et al., 2017. Suitability of Watershed Models to Predict Distributed Hydrologic
1048 Response in the Awramba Watershed in Lake Tana Basin. Land Degradation &
1049 Development, 28(4): 1386-1397. DOI:10.1002/ldr.2608

1050 Monteith, J.L., 1965. Evaporation and the environment, The state and movement of water in
1051 living organisms. 19th Symposia of the Society for Experimental Biology. Cambridge
1052 University Press, London, pp. 205-234.

1053 Morton, D. et al., 2011. Final Report for LCM2007 - the new UK Land Cover Map.

1054 Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R., 2011. Soil and Water Assessment Tool
1055 Theoretical Documentation, Version 2009, Texas Water Resources Institute, Temple.

1056 Ockenden, M.C. et al., 2016. Changing climate and nutrient transfers: Evidence from high
1057 temporal resolution concentration-flow dynamics in headwater catchments. Sci. Total
1058 Environ., 548: 325-339. DOI:10.1016/j.scitotenv.2015.12.086

1059 Ockenden, M.C. et al., 2017. Major agricultural changes required to mitigate phosphorus losses
1060 under climate change. Nature Communications, 8(1): 161. DOI:10.1038/s41467-017-
1061 00232-0

1062 Outram, F.N. et al., 2014. High-frequency monitoring of nitrogen and phosphorus response in
1063 three rural catchments to the end of the 2011–2012 drought in England. Hydrol. Earth
1064 Syst. Sci., 18(9): 3429-3448. DOI:10.5194/hess-18-3429-2014

1065 Overton, D.E., 1966. Muskingum flood routing of upland streamflow. J. Hydrol., 4: 185-200.
1066 DOI:http://dx.doi.org/10.1016/0022-1694(66)90079-5

1067 Owen, G.J. et al., 2012. Monitoring agricultural diffuse pollution through a dense monitoring
1068 network in the River Eden Demonstration Test Catchment, Cumbria, UK. Area, 44(4):
1069 443-453. DOI:10.1111/j.1475-4762.2012.01107.x

1070 Page, T., Beven, K.J., Freer, J., Jenkins, A., 2003. Investigating the Uncertainty in Predicting
1071 Responses to Atmospheric Deposition using the Model of Acidification of

1072 Groundwater in Catchments (MAGIC) within a Generalised Likelihood Uncertainty
1073 Estimation (GLUE) Framework. Water, Air, and Soil Pollution, 142(1): 71-94.
1074 DOI:10.1023/a:1022011520036

1075 Page, T., Beven, K.J., Freer, J., Neal, C., 2007. Modelling the chloride signal at Plynlimon,
1076 Wales, using a modified dynamic TOPMODEL incorporating conservative chemical
1077 mixing (with uncertainty). Hydrological Processes, 21(3): 292-307.
1078 DOI:10.1022/hyp.6186

1079 Page, T., Beven, K.J., Whyatt, D., 2004. Predictive Capability in Estimating Changes in Water
1080 Quality: Long-Term Responses to Atmospheric Deposition. Water, Air, and Soil
1081 Pollution, 151(1): 215-244. DOI:10.1023/B:WATE.0000009893.66091.ec

1082 Page, T., Heathwaite, A.L., Thompson, L.J., Pope, L., Willows, R., 2012. Eliciting fuzzy
1083 distributions from experts for ranking conceptual risk model components. Environ.
1084 Modell. Softw., 36: 19-34. DOI:http://dx.doi.org/10.1016/j.envsoft.2011.03.001

1085 Pappenberger, F. et al., 2006. Influence of uncertain boundary conditions and model structure
1086 on flood inundation predictions. Advances in Water Resources, 29(10): 1430-1449.
1087 DOI:10.1016/j.advwatres.2005.11.012

1088 Perks, M.T. et al., 2015. Dominant mechanisms for the delivery of fine sediment and
1089 phosphorus to fluvial networks draining grassland dominated headwater catchments.
1090 Sci. Total Environ., 523: 178-190. DOI:http://dx.doi.org/10.1016/j.scitotenv.2015.03.008

1091 Radcliffe, D.E., Freer, J., Schoumans, O., 2009. Diffuse Phosphorus Models in the United
1092 States and Europe: Their Usages, Scales, and Uncertainties. J. Environ. Qual., 38(5):
1093 1956-1967. DOI:10.2134/jeq2008.0060

1094 Rankinen, K., Karvonen, T., Butterfield, D., 2006. An application of the GLUE methodology
1095 for estimating the parameters of the INCA-N model. Sci. Total Environ., 365(1-3): 123-
1096 139. DOI:10.1016/j.scitotenv.2006.02.034

1097 Schneiderman, E.M. et al., 2007. Incorporating variable source area hydrology into a curve-
1098 number-based watershed model. Hydrological Processes, 21(25): 3420-3430.
1099 DOI:10.1002/hyp.6556

1100 Schoumans, O.F. et al., 2009. Evaluation of the difference of eight model applications to assess
1101 diffuse annual nutrient losses from agricultural land. Journal of Environmental
1102 Monitoring, 11(3): 540-553. DOI:10.1039/B823240G

1103 Schuol, J., Abbaspour, K.C., 2006. Calibration and uncertainty issues of a hydrological model
1104 (SWAT) applied to West Africa. Adv. Geosci., 9: 137-143. DOI:10.5194/adgeo-9-137-
1105 2006

1106 Shen, Z.Y., Chen, L., Chen, T., 2012a. Analysis of parameter uncertainty in hydrological and
1107 sediment modeling using GLUE method: a case study of SWAT model applied to Three
1108 Gorges Reservoir Region, China. Hydrol. Earth Syst. Sci., 16(1): 121-132.
1109 DOI:10.5194/hess-16-121-2012

1110 Shen, Z.Y., Chen, L., Liao, Q., Liu, R.M., Hong, Q., 2012b. Impact of spatial rainfall variability
1111 on hydrology and nonpoint source pollution modeling. J. Hydrol., 472: 205-215.
1112 DOI:10.1016/j.jhydrol.2012.09.019

1113 Shen, Z.Y., Chen, L., Liao, Q., Liu, R.M., Huang, Q., 2013. A comprehensive study of the
1114 effect of GIS data on hydrology and non-point source pollution modeling. Agricultural
1115 Water Management, 118: 93-102. DOI:10.1016/j.agwat.2012.12.005

1116 Taylor, S.D., He, Y., Hiscock, K.M., 2016. Modelling the impacts of agricultural management
1117 practices on river water quality in Eastern England. Journal of Environmental
1118 Management, 180: 147-163. DOI:http://dx.doi.org/10.1016/j.jenvman.2016.05.002

1119 Team, R.C., 2016. R: A language and Environment for Statistical Computing., R Foundation
1120 for Statistical Computing, https://www.r-project.org/, Vienna, Austria.

1121 van Griensven, A. et al., 2006. A global sensitivity analysis tool for the parameters of multi-
1122 variable catchment models. J. Hydrol., 324(1-4): 10-23.
1123 DOI:10.1016/j.jhydrol.2005.09.008

1124 Vrugt, J.A., Sadegh, M., 2013. Toward diagnostic model calibration and evaluation:
1125 Approximate Bayesian computation. Water Resour. Res., 49(7): 4335-4345.
1126 DOI:10.1002/wrcr.20354

1127 Wang, H., Sun, F., 2016. Impact of LUCC on Streamflow using the SWAT Model over the
1128 Wei River Basin on the Loess Plateau of China. Hydrol. Earth Syst. Sci. Discuss., 2016:
1129 1-30. DOI:10.5194/hess-2016-332

1130 Westerberg, I., Guerrero, J.L., Seibert, J., Beven, K.J., Halldin, S., 2011. Stage-discharge
1131 uncertainty derived with a non-stationary rating curve in the Choluteca River,
1132 Honduras. Hydrological Processes, 25(4): 603-613. DOI:10.1002/hyp.7848

1133 Whitehead, P.G. et al., 2013. A cost-effectiveness analysis of water security and water quality:
1134 impacts of climate and land-use change on the River Thames system. Philosophical
1135 Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences,
1136 371(2002). DOI:10.1098/rsta.2012.0413

Williams, J.R., 1990. The Erosion-Productivity Impact Calculator (EPIC) Model - A Case-History. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 329(1255): 421-428. DOI:10.1098/rstb.1990.0184

Woznicki, S.A., Nejadhashemi, A.P., 2014. Assessing uncertainty in best management practice effectiveness under future climate scenarios. Hydrological Processes, 28(4): 2550-2566. DOI:10.1002/hyp.9804

Yen, H., Hoque, Y., Harmel, R.D., Jeong, J., 2015. The impact of considering uncertainty in measured calibration/validation data during auto-calibration of hydrologic and water quality models. Stoch. Environ. Res. Risk Assess., 29(7): 1891-1901. DOI:10.1007/s00477-015-1047-z

Zhang, P. et al., 2014. Uncertainty of SWAT model at different DEM resolutions in a large mountainous watershed. Water Research, 53: 132-144. DOI:10.1016/j.watres.2014.01.018

Zhang, T. et al., 2013. Estimating phosphorus delivery with its mitigation measures from soil to stream using fuzzy rules. Soil Use and Management, 29: 187-198. DOI:10.1111/j.1475-2743.2012.00433.x

## Figure Captions

**Figure 1: Summary of spatial data in the Newby Beck catchment. Panel a) shows the catchment topography, panel b) shows the locations of the monitoring station (discharge and total phosphorus (TP)), weather station and rain gauges, panel c) shows the main soil classes in the catchment and panel d) shows the broad land use classifications.**

**Figure 2: Generalised likelihood uncertainty estimation (GLUE) likelihood distributions, based upon the evaluation of models using criteria set for all time steps (normalized scores of ± 6.72), of Qsim (simulated discharge), normalised score for Q (discharge), TP loadsim (simulated total phosphorus) and normalised scores for TP, respectively, against observations (panels A-D). The plots are repeated for the low flow periods (panels E-H), rising time-steps (panels I-L), falling (recession) time-steps (panels M-P) and high flow periods (panels Q-T). The areas between the distribution percentiles max/min, $5^{th}$/$95^{th}$ and $25^{th}$/$75^{th}$ are shown in grey shades of increasing intensity. The medians of the distribution are shown by black dots. 1:1 lines and normalised scores of 0 lines have been added for orientation**

**Figure 3: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus loads (b) for Newby Beck outlet (part of the calibration period) based on normalized scores on both discharge and total phosphorus (TP) load evaluation measures when criteria (normalized scores of ± 6.72) set over all model time-steps (1016 simulations). The black line in each plot shows the observed discharge (a) and TP loads (b), respectively. The dashed lines show the uncertainty limits on the calibration data.**

**Figure 4: Generalised Likelihood Uncertainty Estimation (GLUE) likelihood distributions of, based upon the evaluation of models using criteria set for high and low flow periods only (normalized scores of ± 5.30), Qsim (simulated discharge), normalised score for Q (discharge), TP loadsim (simulated total phosphorus) and normalised scores for TP, respectively, against observations (panels A-D). The plots are repeated for the low flow periods (panels E-H), rising time-steps (panels I-L), falling (recession) time-steps (panels M-P) and high flow periods (panels Q-T). The areas between the distribution percentiles max/min, $5^{th}$/$95^{th}$ and $25^{th}$/$75^{th}$ are shown in grey shades of increasing intensity. The medians of the distribution are shown by black dots. 1:1 lines and normalised scores of 0 lines have been added for orientation.**

**Figure 5: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus loads (b) for Newby Beck outlet (part of the calibration period) based on normalized scores on both discharge and total phosphorus (TP) load evaluation measures when criteria (normalized scores of ± 5.30) set over high and low flow time-steps only (1001 simulations). The black line in each plot shows the observed discharge (a) and TP loads (b), respectively. The dashed lines show the uncertainty limits on the calibration data.**

**Figure 6: Dotty plots for 39 of the parameters varied in the Monte-Carlo runs. Parameter names and definitions are shown in Table 1. These are based on the 1016 behavioural simulations evaluated across all time-steps (normalized scores of ± 6.72).**

**Figure 7: Dotty plots for 39 of the parameters varied in the Monte-Carlo runs. Parameter names and definitions are shown in Table 1. These are based on the 1001 behavioural simulations evaluated across the high and low flow time-steps only (normalized scores of ± 5.30).**

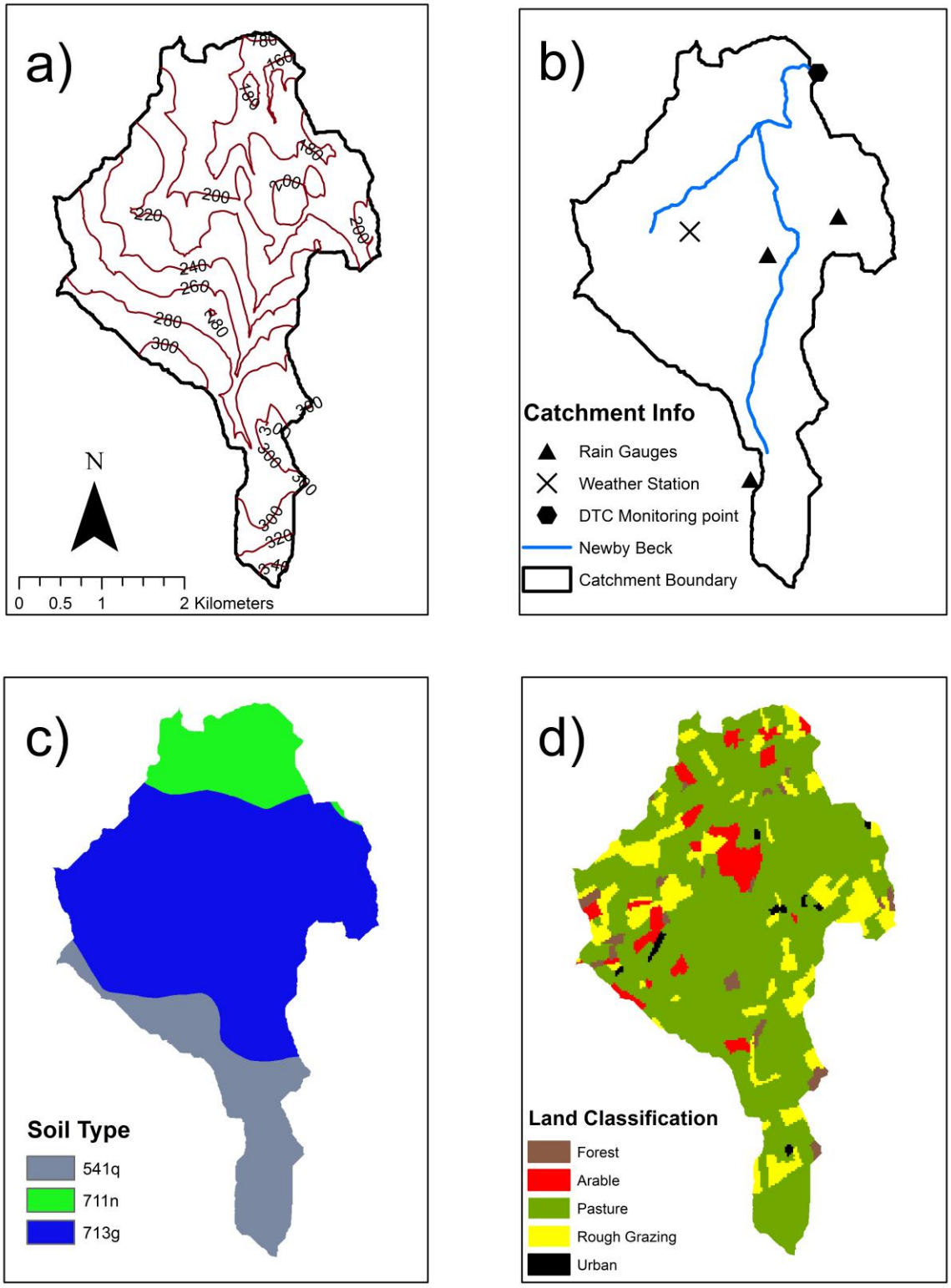
**Figure 8: Breakdown of classification of time-steps resulting in model failure for the 1016 simulations constrained on all time-steps (upper panel) and the 1001 simulations constrained on the high and low flow periods only (lower panel). The bars show the median % contribution to failing time-steps and the error bars show the 2.5/97.5$^{th}$ percentiles from the Generalised Likelihood Uncertainty Estimation (GLUE) weighted distributions.**

**Figure 9: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus (TP) loads (b) for Newby Beck outlet during the validation period (winter of the 2013-2014 Hydrological year) using the 1016 behavioural simulations accepted on both discharge and total phosphorus load criteria when evaluating constrained across all time-steps. The black line in each plot shows the observed discharge (a) and TP loads (b), respectively. The dashed lines show the uncertainty limits on the calibration data.**

**Figure 10: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus (TP) loads (b) for Newby Beck outlet during the validation period (winter of the 2013-2014 Hydrological year) using the 1001 behavioural simulations accepted on both discharge and total phosphorus load criteria when evaluating constrained across high and low flow time-steps only. The black line in each plot shows the observed discharge (a) and TP loads (b), respectively.**

**Figure 11: Generalized Likelihood Uncertainty Estimation (GLUE) weighted prediction bounds (green shading) for discharge (a) and total phosphorus loads (b) for Newby Beck outlet (part of the calibration period) based on normalized scores on both discharge and total phosphorus (TP) load evaluation measures when criteria set over 95% of time steps (1057 simulations). The black line in each plot shows the observed discharge (a) and TP loads (b), respectively.**

1238 **Figure 1:**




1239

1240

**Figure 2:**

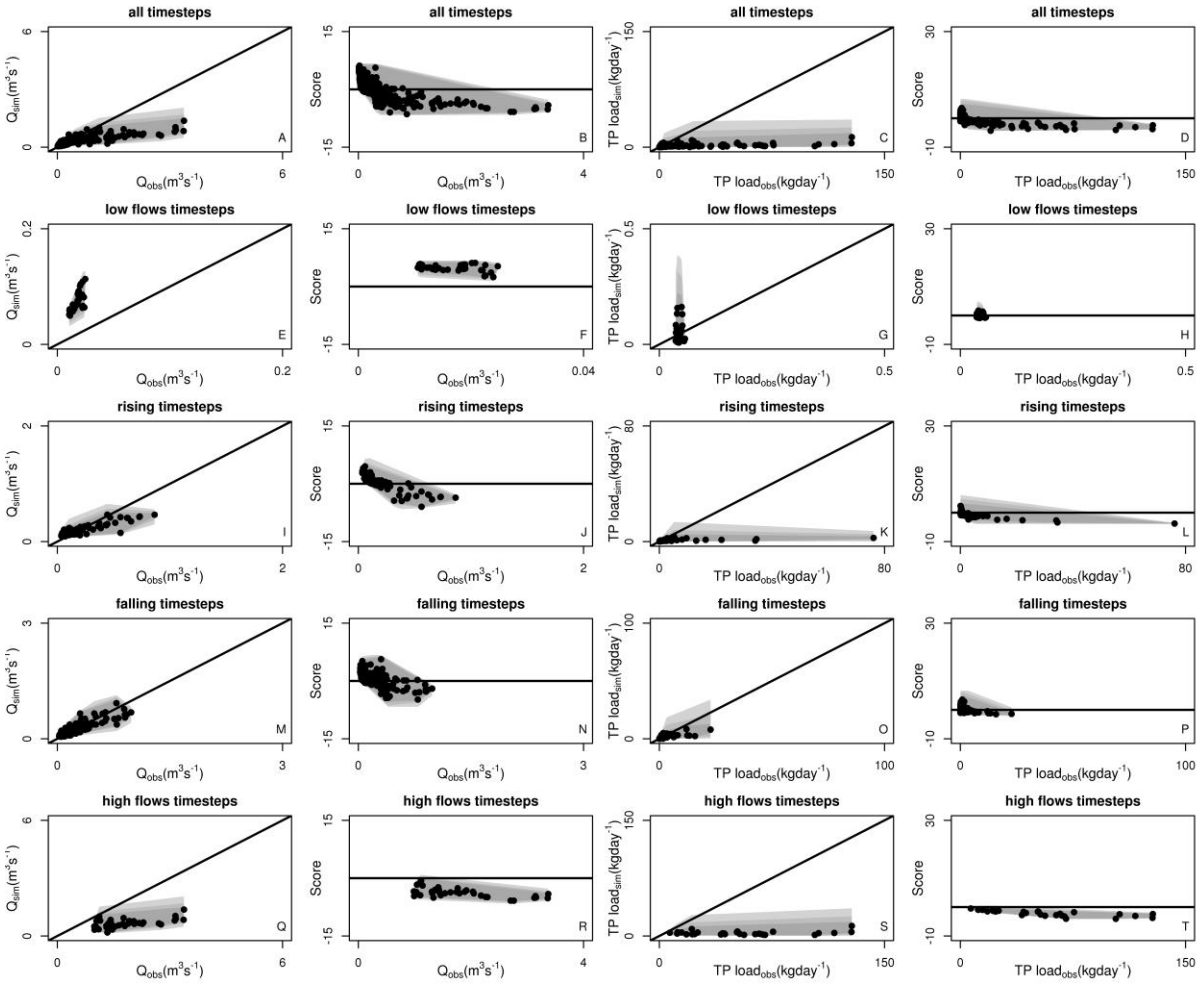

all timesteps
low flows timesteps
rising timesteps
falling timesteps
high flows timesteps
Score
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T

**Figure 3:**

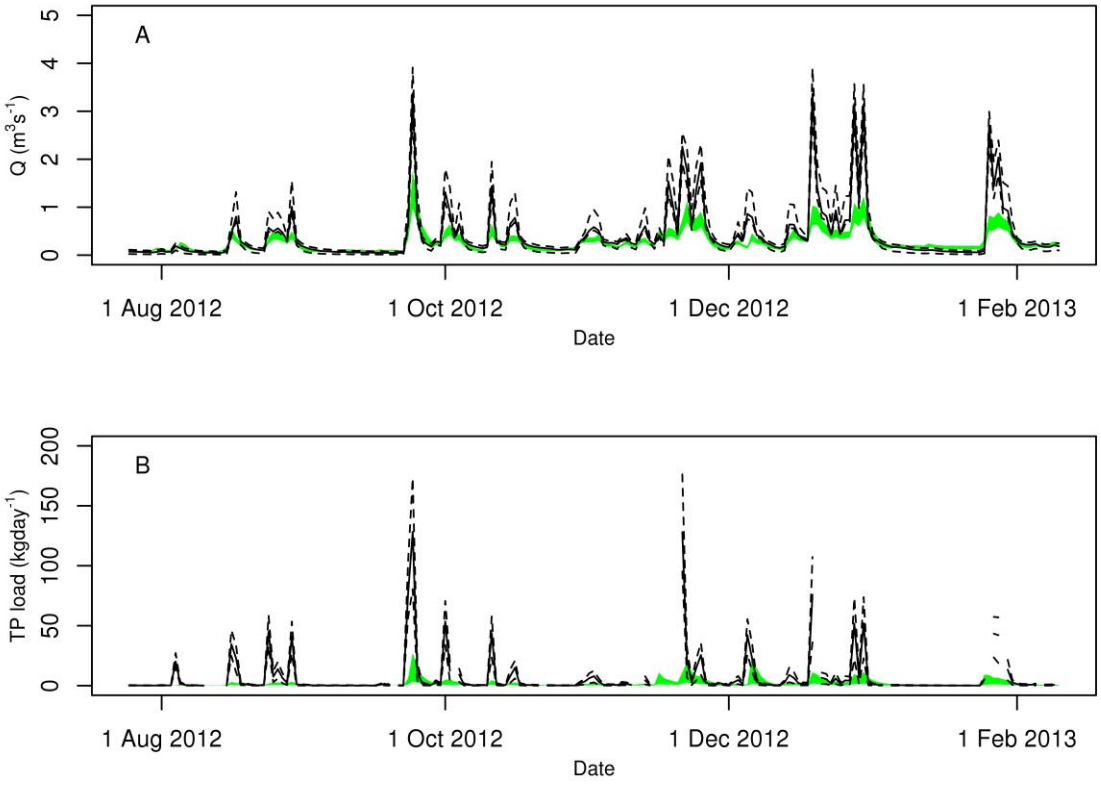

A
Q ($m^3s^{-1}$)
0
1
2
3
4
5
1 Aug 2012
1 Oct 2012
1 Dec 2012
1 Feb 2013
Date
B
TP load ($kgday^{-1}$)
0
50
100
150
200
1 Aug 2012
1 Oct 2012
1 Dec 2012
1 Feb 2013
Date

**Figure 4:**

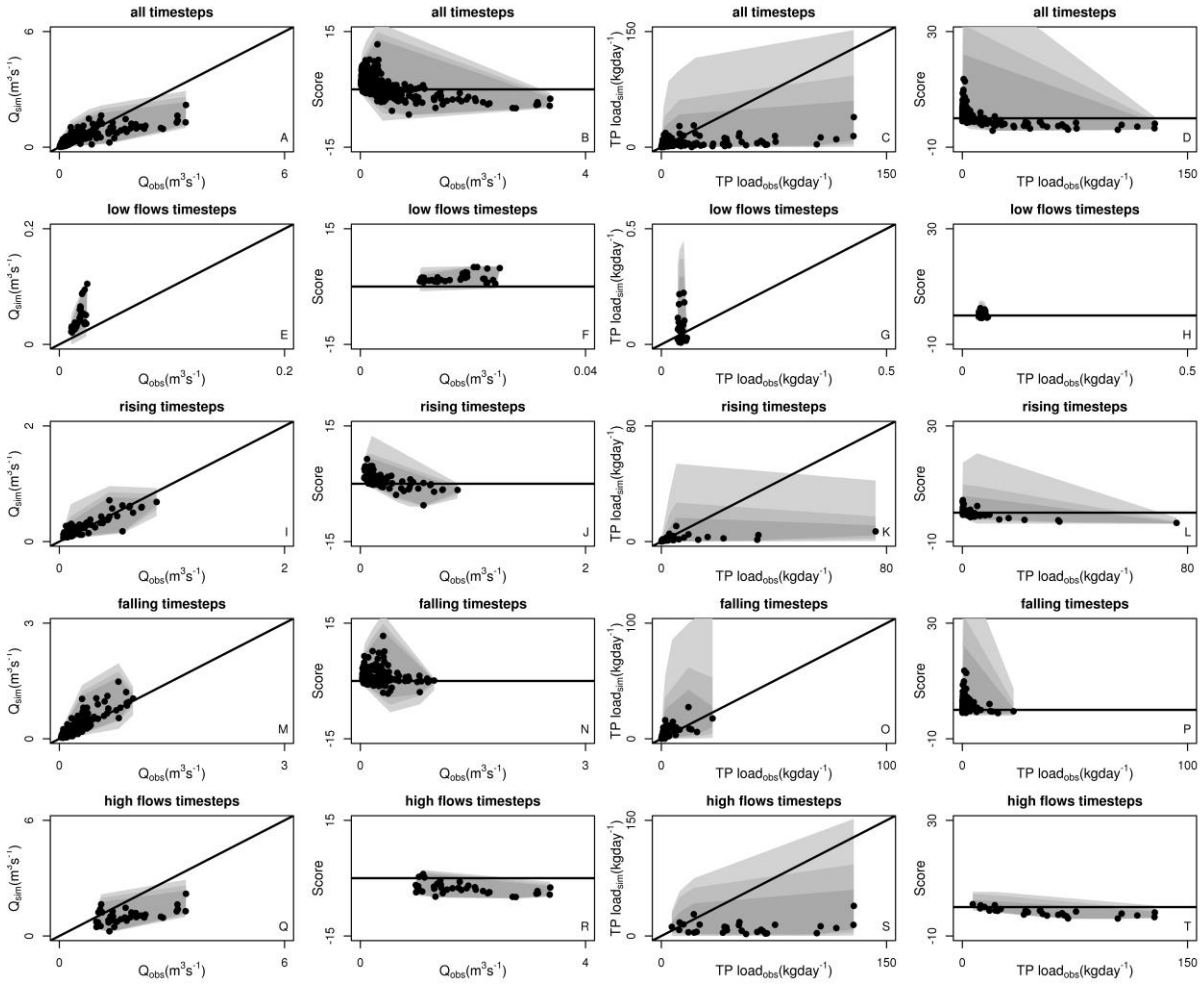

all timesteps
low flows timesteps
rising timesteps
falling timesteps
high flows timesteps
Qsim(m3s-1)
Score
TP loadsim(kgday-1)
Qobs(m3s-1)
TP loadobs(kgday-1)
A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T

**Figure 5:**

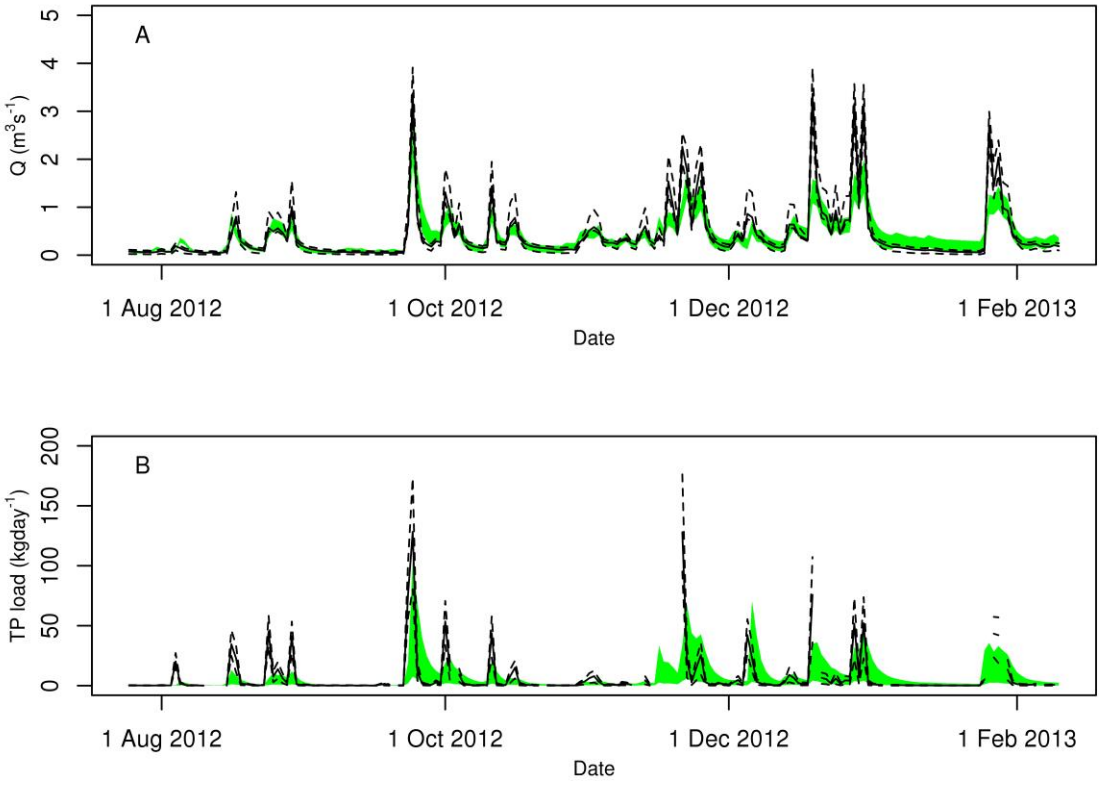

A
Q ($m^3s^{-1}$)
0
1
2
3
4
5
1 Aug 2012
1 Oct 2012
1 Dec 2012
1 Feb 2013
Date
B
TP load ($kgday^{-1}$)
0
50
100
150
200
1 Aug 2012
1 Oct 2012
1 Dec 2012
1 Feb 2013
Date

**Figure 6:**

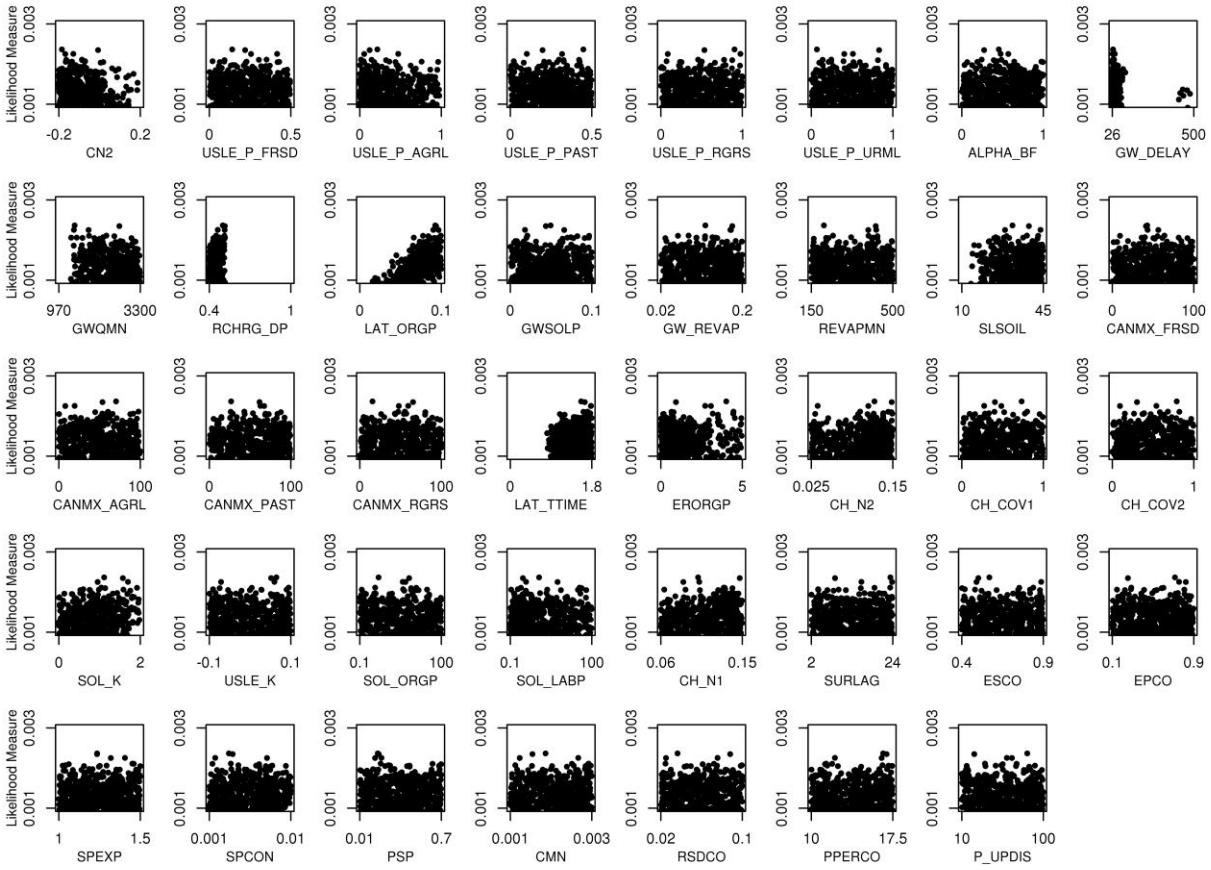

Likelihood Measure
0.001
0.003
-0.2
0.2
CN2
0
0.5
USLE_P_FRSD
0
1
USLE_P_AGRL
0
0.5
USLE_P_PAST
0
1
USLE_P_RGRS
0
1
USLE_P_URML
0
1
ALPHA_BF
26
500
GW_DELAY
970
3300
GWQMN
0.4
1
RCHRG_DP
0
0.1
LAT_ORGP
0
0.1
GWSOLP
0.02
0.2
GW_REVAP
150
500
REVAPMN
10
45
SLSOIL
0
100
CANMX_FRSD
0
100
CANMX_AGRL
0
100
CANMX_PAST
0
100
CANMX_RGRS
0
1.8
LAT_TTIME
0
5
ERORGP
0.025
0.15
CH_N2
0
1
CH_COV1
0
1
CH_COV2
0
2
SOL_K
-0.1
0.1
USLE_K
0.1
100
SOL_ORGP
0.1
100
SOL_LABP
0.06
0.15
CH_N1
2
24
SURLAG
0.4
0.9
ESCO
0.1
0.9
EPCO
1
1.5
SPEXP
0.001
0.01
SPCON
0.01
0.7
PSP
0.001
0.003
CMN
0.02
0.1
RSDCO
10
17.5
PPERCO
10
100
P_UPDIS

**Figure 7:**

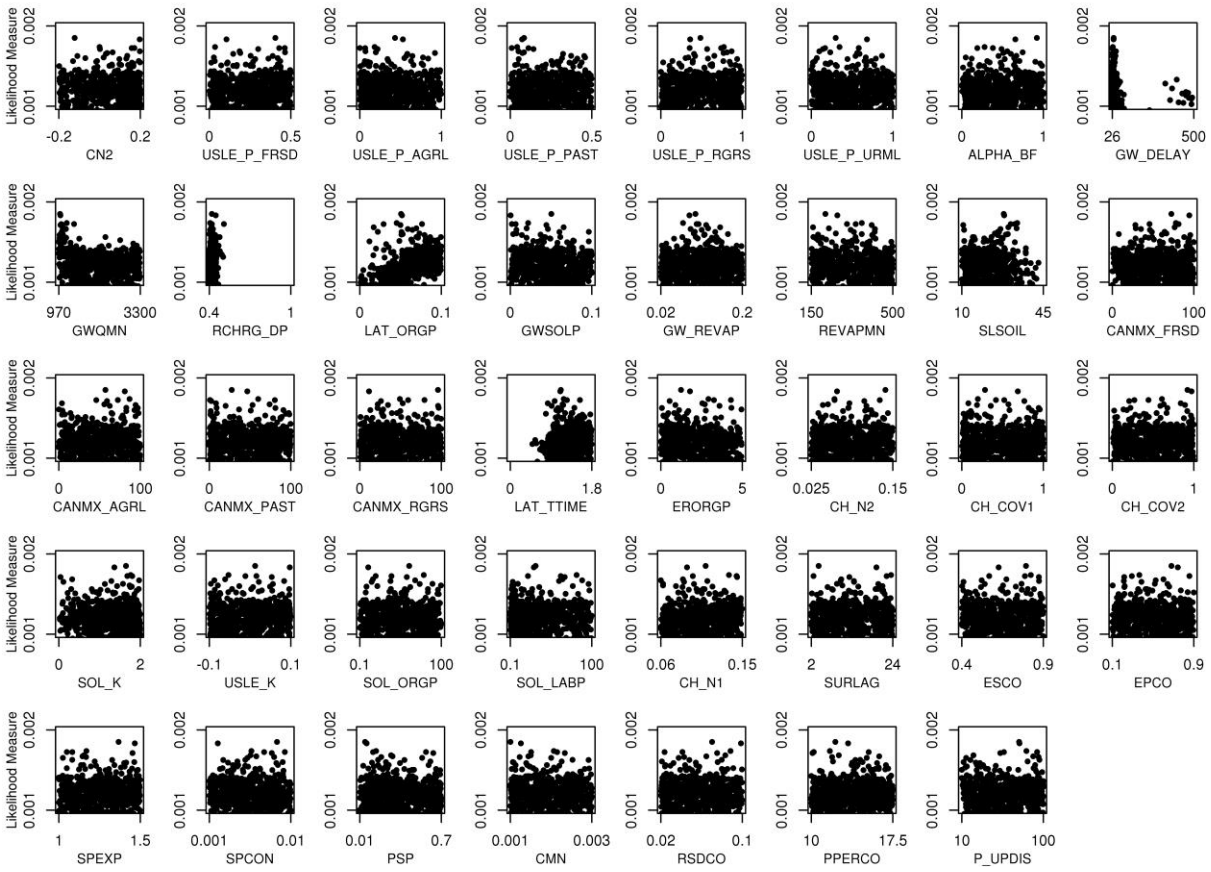

Likelihood Measure
0.001
0.002
CN2
-0.2
0.2
USLE_P_FRSD
0
0.5
USLE_P_AGRL
0
1
USLE_P_PAST
0
0.5
USLE_P_RGRS
0
1
USLE_P_URML
0
1
ALPHA_BF
0
1
GW_DELAY
26
500
GWQMN
970
3300
RCHRG_DP
0.4
1
LAT_ORGP
0
0.1
GWSOLP
0
0.1
GW_REVAP
0.02
0.2
REVAPMN
150
500
SLSOIL
10
45
CANMX_FRSD
0
100
CANMX_AGRL
0
100
CANMX_PAST
0
100
CANMX_RGRS
0
100
LAT_TTIME
0
1.8
ERORGP
0
5
CH_N2
0.025
0.15
CH_COV1
0
1
CH_COV2
0
1
SOL_K
0
2
USLE_K
-0.1
0.1
SOL_ORGP
0.1
100
SOL_LABP
0.1
100
CH_N1
0.06
0.15
SURLAG
2
24
ESCO
0.4
0.9
EPCO
0.1
0.9
SPEXP
1
1.5
SPCON
0.001
0.01
PSP
0.01
0.7
CMN
0.001
0.003
RSDCO
0.02
0.1
PPERCO
10
17.5
P_UPDIS
10
100

**Figure 8:**

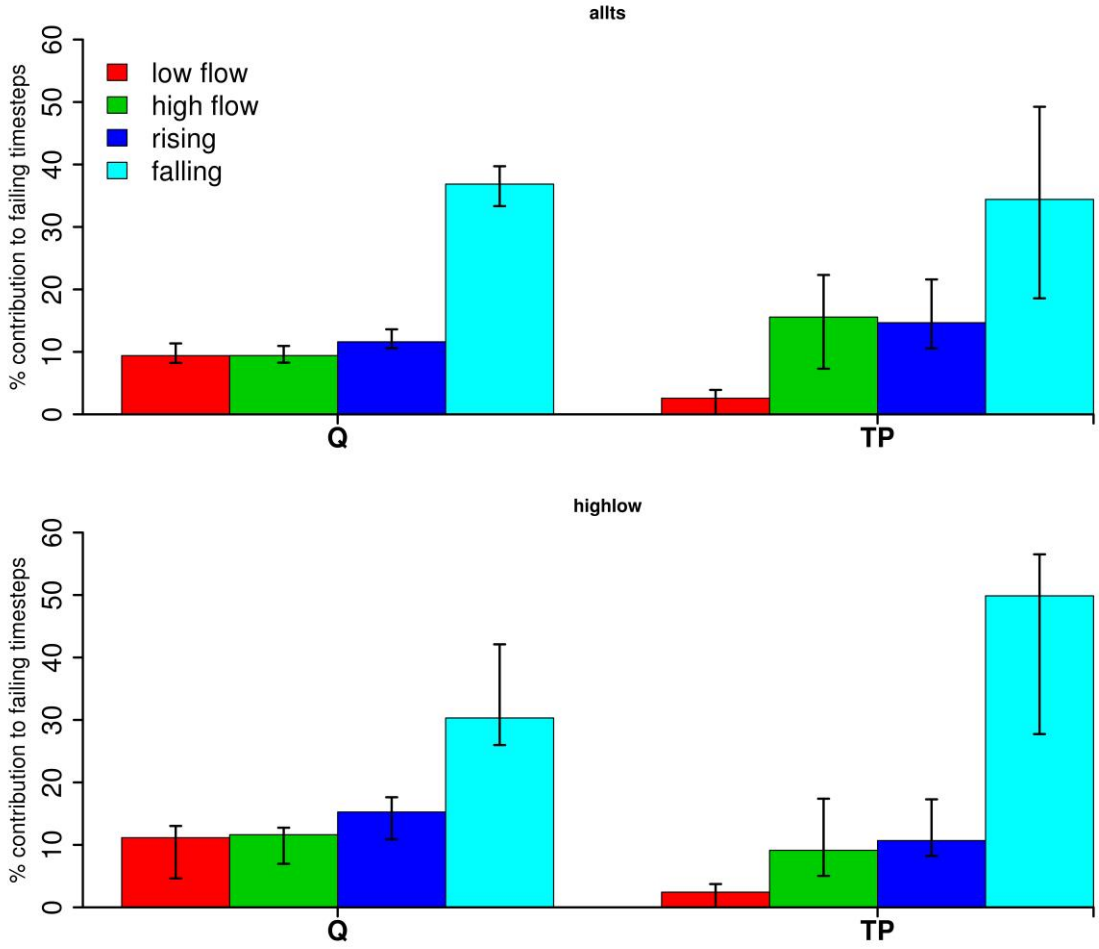

allts
low flow
high flow
rising
falling
% contribution to failing timesteps
0
10
20
30
40
50
60
Q
TP
highlow
% contribution to failing timesteps
0
10
20
30
40
50
60
Q
TP

**Figure 9:**

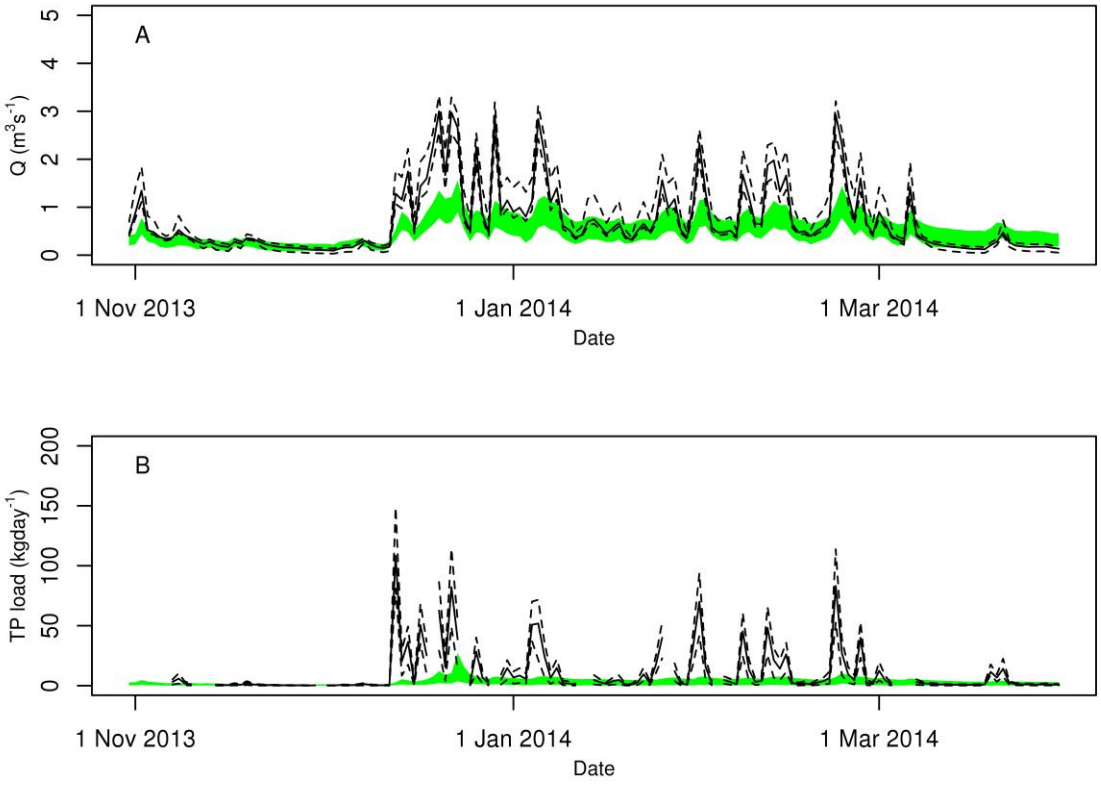

A
Q ($m^3s^{-1}$)
0
1
2
3
4
5
1 Nov 2013
1 Jan 2014
1 Mar 2014
Date
B
TP load ($kgday^{-1}$)
0
50
100
150
200
1 Nov 2013
1 Jan 2014
1 Mar 2014
Date

**Figure 10:**

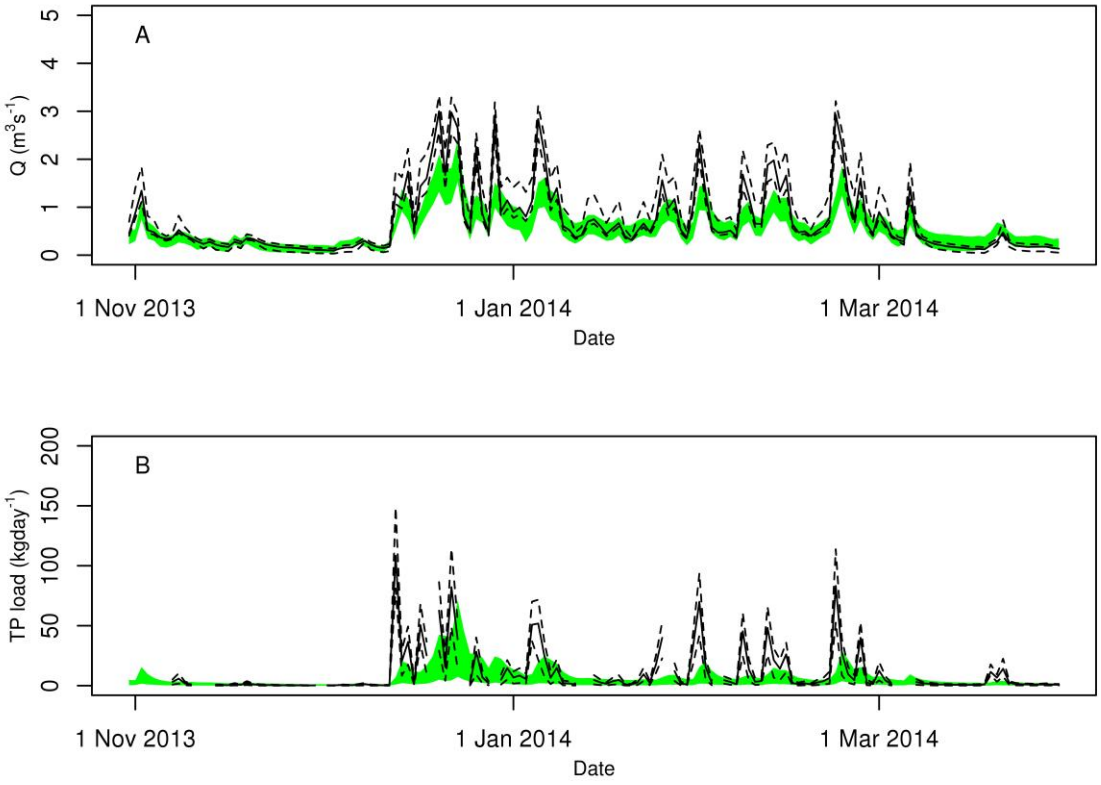
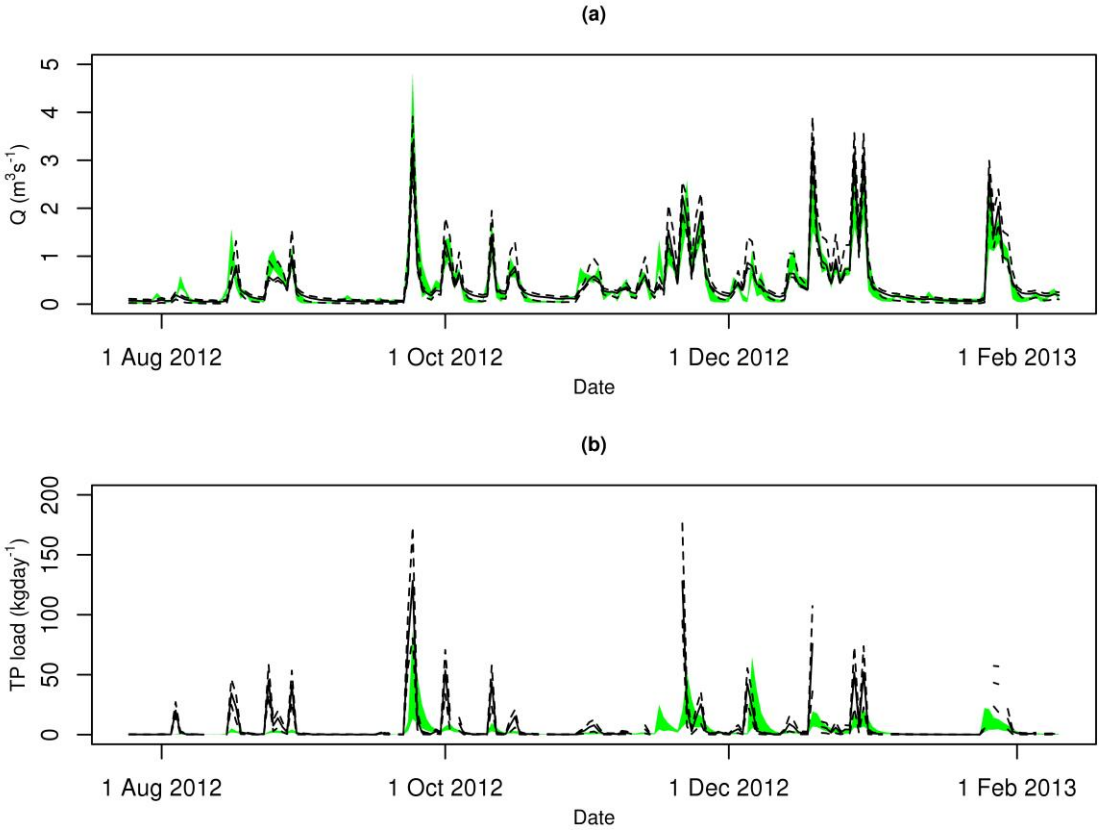

A
Q ($m^3s^{-1}$)
0
1
2
3
4
5
1 Nov 2013
1 Jan 2014
1 Mar 2014
Date
B
TP load ($kgday^{-1}$)
0
50
100
150
200
1 Nov 2013
1 Jan 2014
1 Mar 2014
Date

**Figure 11:**


(a)
Q ($m^3s^{-1}$)
0
1
2
3
4
5
1 Aug 2012
1 Oct 2012
1 Dec 2012
1 Feb 2013
Date
(b)
TP load ($kgday^{-1}$)
0
50
100
150
200
1 Aug 2012
1 Oct 2012
1 Dec 2012
1 Feb 2013
Date

## Highlights

This limits of acceptability approach is applied for the first time to the SWAT model

Identifies exact time steps of poor performance during calibration

Accounts for evaluation data uncertainty in calibration

It may be difficult to obtain sufficient data to drive complex models with confidence