Thanks for your response to the author's reviews and my own. But in fact I am not really yet in agreement with some of the rather quick justifications you have made to the manuscript on the basis of points I have raised so I would like these clarifying further please before I accept this for publication. These issues are important because if under the 'limits of acceptability' the methods are not clear how limits are quantified then in some sense they have no value as part of a considered experimental design.

Thank you for reading our manuscript one more time and for your suggestion to improve it. I realise that I had not understood properly some of the comments in your previous review, hopefully these responses will answer your questions.

Rémi Dupas

1) The authors haven't really justified either the use of using 'prediction intervals', nor that the error assumptions justify the parameteric approach chosen, on the basis of the observed information they have for the rating curve definition. I feel the authors need to deal with these matters better than they have done at the moment in formulation of GLUE LoA or discussion. Furthermore there is no clear rationale as to why a non-linear transformation in rainfall errors (not analysed) would in fact be a surrogate for using this parametric approach having wide uncertainties on the output as some kind of counter balance. I do not think that is well considered as written nor does it deal with the potential differences that might occur if the rainfall errors were . The same goes again for the discharge-concentration uncertainty. What is the proof the parameteric 'prediction interval' error model used relates to the observed error characteristics? – these are both important to get right and/or discuss the limitations/assumptions in them being used!

Response 1:

- In the previous version of the manuscript we justified the choice of a prediction interval rather than a confidence interval based on the fact that some sources of uncertainty were not included (rainfall, DEM, etc.) therefore we chose the largest interval among the two possible ones. This cannot be fully justified unless we analyse error in rainfall (and I do not have good data to do that) and in other sources of error (including those we have not thought about). Now we justify our choice more simply by saying that a prediction interval is an interval in which future observations will likely fall (whereas a confidence interval is an interval in which the mean of repeated observation will likely fall). Because in the TNT2-P model´s evaluation, we want observations to fall in the acceptability interval, a prediction interval is more appropriate.

  Lines 363-367: "A prediction interval is an interval in which future observations will likely fall, while a confidence interval is an interval in which the mean of repeated observation will likely fall. Because in the TNT2-P model´s evaluation, we want each observation to fall in the acceptability interval (section 2.3.3.), a prediction interval was more appropriate."

Of course this justification will only convince the reader if he is convinced that using statistical models was a good choice, which we justified as best as we can in the second point of this response.

- We have added a new discussion paragraph to discuss the drawbacks of using statistical models (three statistical models are used to derive acceptability limits: the rating curve, the SRP concentration uncertainty during baseflow period and the storm event interpolation model).

  Lines 661-671: "Finally, alternative methods to statistical models could be used to derive acceptability limits (in this study three statistical models are used: the rating curve, the SRP concentration uncertainty during baseflow periods and the storm event interpolation model) because statistical models have at least three shortcomings: i) they lump the uncertainty linked to the timing of sampling, the immediate or delayed filtration of the samples, the storage time and the analytical error; ii) the formula chosen adds error to the already existing measurement errors because empirical models are not perfect representation of the system dynamics; iii) they assume a parametric distribution and temporally independent errors which are not always verified in practice. As an alternative, non-parametric methods could be used, but these methods generally require a large number of data points and they are not suitable for extrapolation to extreme values."

- The last criticism in this comment concerns the "What is the proof the parameteric 'prediction interval' error model used relates to the observed error characteristics". A detailed response on the statistical C-Q model is given in comment 3, but we can already say here that we know that the analytical error is an underestimate of the true error in observation (which also includes delayed filtration and analysis) and that the statistical model adds some error related to the extrapolation.

2) I don't understand in the authors response what 'We disagree that the method suggested here is better than ours' is referring to. For a start I am not sure I stated a method was 'better', and secondly it is not at all clear what the context of this response is. So I would like that clarifying please. Perhaps it relates to 1) above…… but then it talks about discharge-concentration curves.

<u>Response 2:</u>

This was a response to the comment "Surely a much more sensible approach…" where it was suggested that we should consider analytical uncertainty rather than a C-Q model to assess uncertainty in SRP load during storm events (if I understood the comment).

The response was in two parts:

- The measurement uncertainty as assessed by the laboratory repetition test is an underestimate of the real uncertainty of autosampler data. The real uncertainty includes, in addition to analytical uncertainty, the issue of samples not immediately filtered and the effect of sample storage.

- We need a statistical model to extrapolate the concentration data from 12h of measurements to a 2-day mean concentration. This model will introduce more error (but this model´s error reflects the missing information originating from the fact that autosampler data did not cover the 2-day period which we use for evaluation).

We added the paragraph:

Lines 403-409: “Two reasons led us to use a statistical model (which also implies the assumption that errors are aleatory and temporally independent): i) the measurement uncertainty as assessed by the laboratory repetition test was an underestimate of the real uncertainty of autosampler data, because it does not include other major sources of error such as delayed filtration and sample decay during storage; ii) it was necessary to extrapolate the sub-daily observation to the daily resolution of the model. The limits of this choice will be discussed in section 4.3.”

Concerning SRP concentration uncertainty during baseflow periods, analytical uncertainty is also an underestimate of the true uncertainty (because other sources of uncertainty such as timing of the grab sampling during the day, or sample storage also play a role), and this was the reason for the use of another statistical model. This was already explained in details in the manuscript.

As acknowledged and discussed in the discussion (see response 1) this choice has several limits which we believe will be solved in part with bankside analyser data, for which observation error will be easier to evaluate.

3) I’m sorry but I am not going to let this issue go of how you derive your load concentration uncertainties and at least make it clearer to the reader what you are doing because at the moment it does not seem consistent or it is certainly not written in a way that makes this year. To be clear from what I can read, you have constructed ‘parametric prediction uncertainty limits’ from the rating curve information. But then you actually do not use these in any way (as far as I can tell) to construct the load uncertainty estimates. You introduce a new model (and a very simple one at that), applied to every storm with a manually applied lag and you gain some very wide uncertainty bounds. Now there are good reasons why in that case the uncertainties will be large, and particularly if that very simple model is not good at describing the dynamics of the discharge-concentration dynamics. If fact as prediction uncertainties it could be argued it is significantly increasing what the potential error limits are in the observations of load. I understand that you need a ‘model’ (although I can still see other ways of doing this) because you wish to extrapolate beyond where you have ISCO samples over days. But that does not mean that you should attempt to be clear exactly what is being done, if that simple model is fit for purpose, the potential issues of increasing load uncertainty estimates over reasonable values if the model is not a description of the system and where you have data if you resampled the expected SRP uncertainties and the discharge uncertainties you have already calculated then what does that look like for the periods you can do this, that finally be clear that you do not seem to be using the uncertainties you have found in discharge to in any way quantify the prediction limits for this simple discharge-

concentration model but instead use standard statistical errors that are yet to be proven. To me this is currently not very clear and not necessarily consistent and it needs to be better explained and discussed……

Response 3:

- The reason why we used statistical models (one for the baseflow periods, one for the storm events) is explained in response 2, and we hope to convince the reader that it was a good choice considering the fact that analytical uncertainty is an underestimate of the true uncertainty and considering the need of extrapolation to the daily resolution of the model. The limits of this choice are now discussed in more details (see response 1).
- The method to derive load acceptability intervals from the 90% prediction interval of discharge and SRP concentration is given in the sentence: "The acceptability limit for daily load was estimated summing up relative uncertainty assessed for discharge and SRP concentration (in percentage)."
  We also had to "combine" the weights for discharge and SRP concentration, both being derived from the statistical model´s error distribution. The method to do this was missing in the manuscript, so added the information:
  Line 458-460: "To "combine" the weights derived from the rating curve and the SRP concentration statistical models, a kernel density estimate (with Gaussian smoothing kernel) was computed to fit 10,000 realisations of the multiplied error models."
- One last critic in this comment concerns the fact that if the C-Q models used to extrapolate SRP during storm events are bad models, the uncertainty interval will inevitably be large. This is true and the reader can make his own opinion on this by looking at the results for each individual model in the supplementary material. We have added a paragraph in the discussion to acknowledge this and to present the perspective that with a bankside analyser (running since April 2016 in this catchment) future work will not require such statistical models because near continuous data will be available and characterization of measurement error will be easier (no difference in the filtration protocol for grab samples and ISCO samples, no delay before analysis and constant analytical error).

Finally the acceptability intervals for storm event loads are also quite large because we stretched the intervals by a factor of 1 -1.6 based on the data we have which show that delayed filtration of autosampler data is causing an apparent loss of SRP.

Lines 424-428: "When comparing autosampler data with data from immediately filtered samples, the ratio obtained had the range 1-1.6 (mean = 1.3), hence autosampler data were underestimates of the true concentration, arguably through adsorption or biological consumption. We used the mean ratio to correct all storm uncertainty intervals by 30% and the range values to extend the upper limit by 60%. "

4) Regarding my minor point 1) noted previously the introduction still states 'In this paper we strive to identify and quantify the different sources of uncertainty in the data when the required quality check

tests have been performed'. Again this needs to be clarified there what the limits of this is in the paper (so only discharge and the P data)

Response 4:

We have added this precision in the introduction: "discharge and SRP concentration data"

5) I do not see how the response to my minor point 5 on the application of homogeneous parameters across the domain has been answered in the response given.

Response 5:

Sorry I had misunderstood this comment.

For the hydrological parameters, we decided to use two soils classes according to the soil map of Curmi et al. (1998) because these authors have measured the hydraulic conductivity for 29 soil cubes in the two soil classes and they appeared to be different (see the following figure extracted from Curmi et al. 1998).

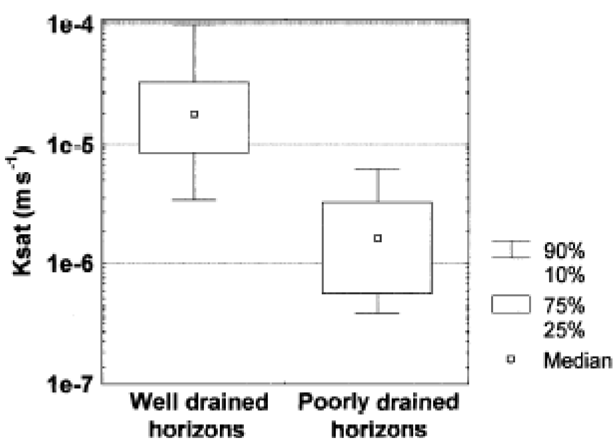

Figure 7. Saturated hydraulic conductivity of the well drained and poorly drained horizons.

We added the sentence:

Lines 383-387: "Experimental determination of saturated hydraulic conductivity (29 soil cores) by Curmi et al. (1998) showed significantly different values for soils classified as well-drained and poorly-drained in the Kervidy-Naizin catchment. The two units were treated as homogeneous, lacking information about the detailed variability in soil hydraulic characteristics at the model grid scale."

For the soil-P model, parameters were considered homogeneous because a previous study in the same catchment showed that the most important factor controlling SRP solubilisation in soils was P Olsen (see section 2.1.3 "Identification of dominant processes") therefore we concentrated our effort on producing a high resolution map of P Olsen (which is an input data to the model) but the parameters to relate this P Olsen to SRP concentration in the soil solution can be considered constant.

We added the sentence:

Line 301-306: "A previous study has shown that soil Olsen P was the most important factor controlling SRP solubilisation in soils of the Kervidy-Naizin catchment (see section 2.1.3.), so other parameters in the soil-P sub-model (section 2.2.2.) were treated as homogeneous in the catchment (the soil classification into well-drained and poorly-drained soils only concerned hydrological parameters)."

6) I think it needs to be justified far better than the response to minor point 6) is somehow justified for such a sparse sample. I'm not going to accept as a scientific evaluation that going from 15,000 – 20,000 simulations 'looked similar' without any justification of what that means. Nor that recognizes that one of the standpoints of using an approach such as GLUE is that the parameter space can be well sampled, or that if a sparse sample must be used there are experimental designs that improve the efficiency of sampling. In effect the authors have a parameter space they are trying to sample that even if they took 2 mid points on each axis this would require 2**30 simulations which is over 1 billion runs. So what convergence would be seen between 15-20K runs! Again the authors appear not to have recognized this at all and the response was not useful in my view and needs to be better justified if they are using GLUE.

Response 6:

We acknowledge 20,000 simulation is a low number and also the fact that the argument that going from 15,000 to 20,000 simulation gave similar results is more a qualitatively appreciation than a real scientific demonstration. We deleted the second part of the sentence (about the 15,000 to 20,000 test) but we maintained the first part where we state that the number of simulations was constrained by computation time.

Several techniques are proposed in the manuscript to solve this problem (some we applied and some we present as perspective):

- First, not all 30 parameters were varied, only 12, and this was already explained so we did not change the paragraph:

  Lines 320-325: "To reduce the number of model runs necessary to explore the parameter space using Monte Carlo simulations, several parameters were given fixed values, or a constant ratio between the two soil types was set (Table 1). In the hydrological sub-model, the parameters to vary were identified in a previous sensitivity analysis (Moreau et al., 2013). In the soil sub-model, all the parameters were varied. Finally, only 12 parameters were varied independently."

- As a perspective (and this was suggested by the reviewer Paul Whitehead), we suggest to use the result of our own sensitivity analysis to vary even less parameters in future applications of the model:

Line 463-465: "This identification of sensitive parameters can be used in future application of the TNT2-P model in the study catchment, as suggested by Whitehead and Hornberger (1984) and Wade et al. (2002b)."

- Also as a perspective we suggest a method to reduce computation time by introducing the concept of hydrological and chemical similarity. The following paragraph was extended to address this comment (additional sentences are underlined):

  Lines 593-603: "It would be interesting to test to what extent moving from an aggregative model with fully distributed information to a semi-distributed model would degrade the model performance while reducing computational cost. This could be achieved by grouping cells according to a hydrological similarity criterion like in Dynamic Topmodel (Beven and Freer, 2001b; Metcalfe et al., 2015) and do the same for similarity in soil P content. Reducing computation time is critical in the context of a GLUE analysis because this method requires the parameter space to be sampled adequately to identify those models to be considered acceptable. This is debatable here because 12 parameters were varied and only 20,000 model runs were performed. It is therefore possible that some regions of the parameter space with acceptable models might not have been sampled."

7) Similar issues of not really providing a useful response go with the response to minor point 4) and 5). First there is still seemingly no analyses of why 20m DEM resolution is needed that is explicitly written in the model setup, so if somehow the hillslope characterization is being lost if the resolution was lower then in what way is some critical threshold being reached for the D8 sharing downslope? How has that been confirmed given the simplifications in general in the model? I still don't see how this all squares with the authors own statement that the main SRP transportation processes are controlled hydrologically by valley bottom groundwater fluctuations (between page 6-7).

Response 7:

We have added the argument:

Lines 307-314: "A 20 m resolution was chosen for the DEM and the soil Olsen P raster map to allow a detailed representation of the interaction of the groundwater table (as simulated by the hydrological model) and the soil Olsen P (as given by the soil Olsen P map). Indeed the soil saturation and soil Olsen P can be very different in a narrow zone close to the stream compared to upslope due to the presence of a 5 to 50 m unfertilized buffer zone with lower Olsen P compared to fertilized fields. The Olsen P value close to the stream has a determining influence on SRP transfer, because this area is the most frequently connected to the stream, so a coarser resolution of the raster maps would degrade representation of the system."

Similarly to the criticism on the number of simulation and the number of soil hydrological classes, the only way to demonstrate that 20m resolution is really important would be to make a formal sensitivity analysis, which we did not do because i) we had already some expert knowledge on the best resolution (see the references about old applications of TOPMODEL in the catchment Bruneau et al., 1995; Franks et al, 1998 and all the TNT2 papers), the dominant processes to include, etc and ii) we were already constrained by calculation times to test all the different alternative possibilities.

Regarding minor point 5) here is nothing in the additional sentence added that at all discusses how these parameters are homogeneous across the catchment to the level they have been applied. No evidence is provided to say why that is realistic in the fully distributed model design or why 2 classes are the dominant hydrological-chemical classifications. This again needs to be improved and the responses were quite weak.

<u>Response 8:</u>

I have understood the criticism now and additional justification is given in response 5.

Uncertainty assessment of a dominant-process catchment model of dissolved phosphorus transfer

R. Dupas[1], J. Salmon-Monviola[1], K. Beven[2], P. Durand[1], P.M. Haygarth[2], M.J. Hollaway[2], C. Gascuel-Odoux[1]

[1] INRA, Agrocampus Ouest, UMR1069 SAS, F-35000 Rennes, France

[2] Lancaster Environment Centre, Lancaster University, Lancaster, United Kingdom, LA1 4YQ

Correspondence to: R. Dupas (remi.dpas@gmail.com)

Abstract

We developed a parsimonious topography-based hydrologic model coupled with a soil biogeochemistry sub-model in order to improve understanding and prediction of Soluble Reactive Phosphorus (SRP) transfer in agricultural headwater catchments. The model structure aims to capture the dominant hydrological and biogeochemical processes identified from multiscale observations in a research catchment (Kervidy-Naizin, 5 km²). Groundwater fluctuations, responsible for the connection of soil SRP production zones to the stream, were simulated with a fully-distributed hydrologic model at 20 m resolution. The spatial variability of the soil phosphorus content~~status~~ and the temporal variability of soil moisture and temperature, which had previously been identified as key controlling factors of SRP solubilisation in soils, were included as part of an empirical soil biogeochemistry sub-model. The modelling approach included an analysis of the information contained in the calibration data and propagation of uncertainty in model predictions using a GLUE "limits of acceptability" framework. Overall, the model appeared to perform well given the uncertainty in the observational data, with a Nash-Sutcliffe efficiency on daily SRP loads between 0.1 and 0.8 for acceptable models. The role of hydrological connectivity via groundwater fluctuation, and the role of increased SRP solubilisation following dry/hot periods were captured well. We conclude that in the absence of near continuous monitoring, the amount of information contained in the data is limited hence parsimonious models are more relevant than highly parameterised models. An analysis of uncertainty in the data is recommended for model calibration in order to provide reliable predictions.

1 Introduction

Excessive phosphorus (P) concentrations in freshwater bodies result in increased eutrophication risk worldwide (Carpenter et al., 1998; Schindler et al., 2008). Eutrophication restricts economic use of water and poses a serious ~~health~~ hazard to ecosystems and humans~~,~~ ~~due to the potential development of harmful cyanobacteria (Bradley et al., 2013;~~ (Serrano et al., 2015). In western countries, reduction of point source P emissions in the last two decades has resulted in a proportionally increasing contribution of diffuse sources, mainly from agricultural origin (Alexander et al., 2008; Grizzetti et al., 2012; Dupas et al., 2015a). Of particular concern are dissolved P forms, often measured as Soluble Reactive Phosphorus (SRP), because they are highly bioavailable and therefore a likely contributor to eutrophication.

To reduce SRP transfer from agricultural soils it is important to identify the spatial origin of P sources in agricultural landscapes, the biogeochemical mechanisms causing SRP solubilisation in soils~~–and~~ and ~~–~~the dominant transfer pathways, as well as the potential P resorption during transit.~~.~~ Research catchments provide useful data to investigate SRP transport mechanisms: typically, the temporal variations in water quality parameters at the outlet, together with hydroclimatic variables, are investigated to infer spatial origin and dominant transfer pathways of SRP (Haygarth et al., 2012; Outram et al., 2014; Dupas et al., 2015b; Mellander et al., 2015; Perks et al., 2015). Hypotheses drawn from analysis of water quality time series can be further investigated through hillslope monitoring and/or laboratory experiments (Heathwaite and Dils, 2000; Siwek et al., 2013; Dupas et al., 2015c). When dominant processes are considered reasonably known, it is possible to develop computer models, for two main purposes: first, to validate scientific conceptual models, by testing whether model predictions can produce reasonable simulations compared to observations. Of particular interest is the possibility ~~to~~ of testing the capability of a computer model to upscale P processes observed at fine spatial resolution (soil column, hillslope) to a whole catchment. Secondly, if the models survive such validation tests, then they can be useful tools to simulate the response of a catchment system to a future perturbation such as changes in agricultural management and climate changes.

However, process-based P models generally perform poorly compared to, for example, nitrogen models (Wade et al., 2002; Dean et al., 2009; Jackson-Blake et al., 2015a). This is of major concern because poor model performance suggests poor knowledge of dominant processes at the catchment scale, and poor reliability of the modelling tools used to support management. The origin of poor model performance might be conceptual misrepresentations, structural imperfection, calibration problems, irrelevant model evaluation criteria and difficulties in properly assessing the information content of the available data when it is subject to epistemic error. All five causes of poor model performance are intertwined, e.g. model calibration strategy depends on model performance evaluation criteria, which depend on the way the information contained in the observation data is assessed (Beven and Smith, 2015).

A key issue in environmental modelling is the level of complexity one should seek to incorporate in a model structure. Several existing P transfer models, such as INCA (Wade et al., 2002), SWAT (Arnold et al., 1998) and HYPE (Lindstrom et al., 2010) seek to simulate many processes, with the view that complex models are necessary to understand processes and to predict the likely consequences of land-use or climate changes. However, these complex models include many parameters that need to be calibrated, while the amount of data available for calibration is often low. An imbalance between calibration requirement and the amount of available observation data can lead to equifinality issues, i.e. when many model structures or parameter sets lead to acceptable simulation results (Beven, 2006). A consequence of equifinality is the risk of unreliable prediction when an “optimal” set of parameters is used (Kirchner, 2006), and large uncertainty intervals when Monte Carlo simulations are performed (Dean et al., 2009). In this situation, it will be worth exploring parsimonious models that aim to capture the dominant hydrological and biogeochemical processes controlling SRP transfer in agricultural catchment. For example, Hahn et al. (2013) used a soil-type based rainfall-runoff model (Lazzarotto et al., 2006) combined with an empirical model of soil SRP release derived from rainfall simulation experiments over soils with different P content and manure application level/timing (Hahn et al., 2012) to simulate daily SRP load from critical sources areas.

A second key issue, linked to the question of model complexity, concerns model calibration and evaluation. Both calibration and evaluation require assessing the fit of model outputs with observation data. However, observation data are generally not directly comparable with model outputs, because of incommensurability issues and/or because they contain errors (Beven, 2006; 2009). Typically, predicted daily concentrations and/or loads are evaluated against data from grab samples collected on a daily or weekly basis. The information content of these data must be carefully evaluated to propagate uncertainty in the data into model predictions (Krueger et al., 2012). Uncertainty in grab sample data might stem from i) sampling frequency problems and ii) measurement problems (Lloyd et al., ~~2015~~2016). Grab sample data represent a specific point in the stream cross-section, which can differ from the cross section mean concentration (Rode and Suhr, 2007), and a snapshot of the concentration at a given time of the day, which can differ from the flow weighted mean daily concentration (McMillan et al. 2012). This difference between observation data and simulation output can be large during storm events in small agricultural catchments, as P concentrations can vary by several orders of magnitudes during the same day (Heathwaite and Dils, 2000; Sharpley et al., 2008). Model evaluation can be severely penalised by this difference, because many popular evaluation criteria such as the Nash-Sutcliffe efficiency (NSE) are sensitive to extreme values and errors in timing (Moriasi et al., 2007). During baseflow periods, it is more likely that grab sample data are comparable to flow-weighted mean daily concentrations, as concentrations vary little during the day and they are usually low in the absence of point sources. However, measurement errors are expected to occur at low concentrations, either due to too long storage times or laboratory imprecision when concentrations come close to detection/quantification limits (Jarvie et al., 2002; Moore and Locke, 2013). Uncertainty in the data can also relate to discharge measurement and input data (e.g. maps of soil P content and rainfall data). In this paper we strive to identify and quantify the different sources of uncertainty in the data when the required quality check tests have been performed (on the discharge and SRP concentration data). A Generalised Likelihood Uncertainty Estimation (GLUE) “limits of acceptability” approach (Beven, 2006; Beven and Smith, 2015) is used to calibrate/evaluate the model.

This paper presents a dominant-process model that couples a topography-based hydrologic model with a soil biogeochemistry sub-model able to simulate daily discharge and SRP loads. The dominant processes included in the hydrologic and soil biogeochemistry sub-models have been identified in previous analyses of multiscale observational data, which have demonstrated on the one hand the control of groundwater fluctuation on connecting soil SRP production zones to the stream (Haygarth et al., 2012; Jordan et al., 2012; Dupas et al., 2015b; 2015d; Mellander et al., 2015), and on the other hand the role of antecedent soil moisture and temperature conditions on SRP solubilisation in soils (Turner and Haygarth, 2001; Blackwell et al., 2009; Dupas et al., 2015c). Model development and application were~~as~~ performed in the Kervidy-Naizin catchment in western France with the objectives of: i) testing if the model was capable of capturing daily variation of SRP load, thus confirming hypotheses on dominant processes; ii) develop a methodology to analyse and propagate uncertainty in the data into model prediction using a “limits of acceptability” approach. ~~Model development and analysis of uncertainty in the data are interlinked in this approach.~~

2 Material and methods

2.1 Study catchment

2.1.1 Site description

Kervidy–Naizin is a small (4.94 km²) agricultural catchment located in central Brittany, Western France (48°N, 3°W). It belongs to the AgrHyS environmental research observatory (http://www6.inra.fr/ore_agrhys_eng), which studies the impact of agricultural activities and climate change on water quality (Molenat et al., 2008; Aubert et al., 2013; Salmon-Monviola et al., 2013~~; Humbert et al., 2014~~). The catchment (Fig. 1) is drained by a stream of second Strahler order, which generally dries up in August and September. The climate is temperate oceanic, with mean ± standard deviations of annual cumulative precipitation and specific discharge ~~averaging~~ of 854 ± 179 mm and 290 ± 106 mm, respectively, from 2000 to 2014. Mean annual ± standard deviation of temperature is 11.2 ± 0.6°C. Elevation ranges from 93 to 135 m above sea level. Topography is gentle, with maximum slopes not exceeding 5%. The bedrock consists of impervious, locally fractured Brioverian schists and is capped by several metres of unconsolidated weathered material and silty, loamy soils. The hydrological behaviour is dominated by the development of a water table that varies seasonally along the hillslope. In the upland domain, consisting of well drained soils, the water table remains below the soil surface throughout the year, varying in depth from 1 to > 8 m. In the wetland domain, developed near the stream and consisting of hydromorphic soils, the water table is shallower, remaining near the soil surface generally from October to April each year. The land use is mostly agriculture, specifically arable crops and confined animal production (dairy cows and pigs). A farm survey conducted in 2013 led to the following land use subdivisions: 35% cereal crops, 36% maize, 16% grassland and 13% other crops (rape seed, vegetables). Animal density was estimated as high as 13 livestock units ha^{-1} in 2010. Estimated soil P surplus was~~is~~ 13.1 kg P ha^{-1} yr^{-1} (Dupas et al., 2015b) and soil extractable P in 2013 (Olsen et al., 1954) wa~~i~~s 59 ± 31 mg P kg^{-1} (n = 89 samples). A survey targeting riparian areas highlighted the legacy of high soil P content in these currently unfertilized areas (Dupas et al., 2015c). No point source emissions we~~a~~re recorded but scattered dwellings with septic tanks we~~a~~re present in the catchment.

2.1.2 Hydroclimatic and chemical monitoring

Kervidy-Naizin was equipped with a weather station (Cimel Enerco 516i) located 1.1 km from the catchment outlet. It recorded hourly precipitation, air and soil temperatures, air humidity, global radiation, wind direction and speed, that are used to~~and~~ estimate~~s~~ Penman evapotranspiration. Stream discharge was estimated at the outlet with a rating curve and stage measurements from a float-operator sensor (Thalimèdes OTT) upstream of a rectangular weir.

To record both seasonal and within storm dynamics in P concentration, two monitoring strategies complemented each other from October 2013 to August 2015: a daily manual grab sampling at approximately the same time (between 16:00 – 18:00 local time) and automatic high frequency sampling during 14 storm events (autosampler ISCO 6712 Full-Size Portable Sampler, 24 one litre bottles filled every 30 min). The water samples were filtered on-site, immediately after grab sampling and after 1-2 days in the case of autosampling. They were analysed for SRP (ISO 15681) within a fortnight. To assess uncertainty in daily SRP concentration related to sampling time, storage and measurement errors, a second grab sample was taken at a different time of the day (between 11:00 – 15:00 local time) in 36 instances during the study period. The second sample was analysed within 24h with the same method; this second dataset is referred to as verification dataset, as opposed to the reference dataset. Among the 36 pairs of comparable daily samples, 12 were taken during storm events and 24 during baseflow periods. To assess uncertainty in high frequency SRP concentration during storm events due to delayed filtration of autosampler bottles, 5 grab samples were taken during the course of 4 distinct storms and were filtered immediately. The same lab procedure was used to analyse SRP.

2.1.3 Identification of dominant processes from multiscale observations

Observations in the Kervidy-Naizin catchment have highlighted that the temporal variability in stream SRP concentrations could not be related to the calendar of agricultural practices, but rather to hydrological and biogeochemical processes (Dupas et al., 2015b). The primary control of hydrology on SRP transfer has also been evidenced in several other small agricultural catchments (e.g. Haygarth et al, 2012; Jordan et al., 2012; Mellander et al., 2015). In the Kervidy-Naizin catchment, the groundwater fluctuation~~s~~ in valley bottom areas was identified as the main driving factor of SRP transfer, through the hydrological connectivity it creates when the saturated zone~~it~~ intercepts shallow soil layers (Dupas et al., 2015b).

In-situ monitoring of soil pore water at 4 sites (15 cm and 50 cm depths) in the Kervidy-Naizin catchment has shown that mean SRP concentration in soils ~~was~~ is a linear function of Olsen P (Olsen et al., 1954). This reflects current knowledge that a soil P test, or alternatively estimation of a degree of P saturation, can be used to assess solubilisation in soils (Beauchemin and Simard, 1999; McDowell et al., 2002; Schoumans et al., 2015). This linear relationship derived from the data contrasts however with other studies, where threshold values above which SRP solubilisation increases greatly have been identified (Heckrath et al., 1995; Maguire et al., 2002).

Soluble Reactive Phosphorus solubilisation in soil varies seasonally according to antecedent conditions of temperature and soil moisture. Dry and/or hot conditions are favourable to the accumulation of mobile P forms in soils, while water saturated conditions lead to their flushing (Turner et al., 2001; Blackwell et al., 2009; Dupas et al., 2015c).

2.2 Description of the Topography-based Nutrient Transfer and Transformation – Phosphorus model (TNT2-P)

TNT2 was originally developed as a process-based and spatially explicit model simulating water and nitrogen fluxes at a daily time step (Beaujouan et al., 2002) in meso-scale catchments (< 50 km^2). TNT2-N has been widely used for operational objectives, to test the effect of mitigation options proposed by local stakeholders or public policy-makers (Moreau et al., 2012; Durand et al., 2015), on nitrate fluxes and concentrations in rivers.

TNT2-P uses a modified version of the hydrological sub-model in TNT2-N, to which a P biogeochemistry sub-model was added to simulate SRP solubilisation in soils.

2.2.1 Hydrological sub-model

The assumptions in the hydrological sub-model are derived from TOPMODEL which has previously been applied to the Kervidy-Naizin catchment (Bruneau et al., 1995; Franks et al., 1998): 1) the effective hydraulic gradient of the saturated zone is approximated by the local topographic surface gradient (tan β). It is calculated in each cell of a Digital Elevation Model (DEM) at the beginning of the simulation; 2) the effective downslope transmissivity (parameter T) of the soil profile in each cell of the DEM is a function of the soil moisture deficit (Sd). Hydraulic conductivity is assumed to decrease~~s~~ exponentially with depth (parameter m, Fig. 2). Hence water fluxes (q) are computed as:

$$q = T * tan\beta * \exp(-\frac{Sd}{m}) \qquad (1)$$

Based on these assumptions, TNT2 computes an explicit cell-to-cell routing of fluxes, using a D8 algorithm. ~~This explicit cell-to-cell routing of fluxes increases computation times compared to TOPMODEL, for which calculations are grouped according to a distribution of hydrologically similar points, but it allows taking account of spatial interactions between soil and groundwater, which has been shown to improve representation of nutrients fluxes and transformations (Beaujouan et al., 2002).~~

To simulate SRP fluxes, ~~the only modification to~~ the hydrological sub-model is used ~~aimed~~ to compute water fluxes from each soil layer by integrating [1] between the maximum depth of the soil layer considered and either:

- estimated groundwater level, if the groundwater table is within the soil layer considered

or

- the minimum depth of the soil layer considered, if the groundwater table above the soil layer considered

In this application of the TNT2-P model, 5 soil layers with a thickness of 10 cm are considered. Hence, 7 flow components are computed in the model:

- overland flow on any saturated surfaces
- 5 sub-surface flow components, one for each soil layer
- deep flow, i.e. flow below the 5 soil layers

2.2.2 Soil-P sub-model

The soil-P sub-model is empirically derived from soil pore water monitoring data (Dupas et al., 2015c), specifically assuming that:

- background SRP concentration in the soil pore water of a given layer is proportional to soil Olsen P;
- seasonal increases in P availability compared to background conditions are determined by biogeochemical processes, controlled by antecedent temperature and soil moisture.

Data show that SRP availability in the soil pore water increases following periods of dry and hot conditions (Dupas et al., 2015c).

Hence, SRP transfer is modelled with parameters that describe both mobilisation and transfer to the stream. A different parameter is used to simulate transfer via overland flow and sub-surface flow.

$$F_{SRP\ overland} = Coef_{SRP\ overland} * P_{Olsen} * q_{overland} \qquad (2)$$

$$F_{SRP\ sub-surface} = Coef_{SRP\ sub-surface} * P_{Olsen} * q_{sub-surface} \qquad (3)$$

Where $F_{SRP\ overland}$ and $F_{SRP\ sub-surface}$ are SRP transfer via overland flow and sub-surface flow for a given soil layer respectively, $q_{overland}$ and $q_{sub-surface}$ are water flows from the same pathways. $\mathrm{Coef}_{\mathrm{SRP\ overland}}$ and $\mathrm{Coef}_{\mathrm{SRP\ sub-surface}}$ are coefficients which vary according to antecedent temperature and soil moisture conditions, such as:

$$Coef_{SRP} = Coef_{background} * (1 + F_T * F_S) \qquad (4)$$

Where $Coef_{SRP}$ is either $\mathrm{Coef}_{\mathrm{SRP\ overland}}$ or $\mathrm{Coef}_{\mathrm{SRP\ sub-surface}}$, and F_T and F_S are temperature and soil moisture factors, respectively. F_T and F_S are expressed as:

$$F_T = \exp\left(\frac{mean(temperature, i\ days) - T1}{T2}\right) \qquad (5)$$

$$F_S = 1 - \left(\frac{\mathrm{mean(water\ concentent, i\ days\ )}}{maximum\ water\ content}\right)^{S1} \qquad (6)$$

Where T1, T2 and S1 are parameters to be calibrated ~~coefficients~~. The antecedent condition time length consists in a period of i=100 days. Both soil temperature and soil moisture are estimated by the TNT2 soil module (Moreau et al., 2013). Because soil moisture in the deep soil layers can differ significantly from that of shallow soil layers, two values of F_S are calculated for two soil depth ranges 0-20 cm and 20-50 cm. The temperature factor F_T was calculated as an average value for the entire 0-50 cm soil profile ~~0-50 cm~~. Contrary to the water fluxes, SRP fluxes are not routed cell-to-cell, because we lack~~ed~~ knowledge of the rate of SRP re-adsorption in downslope cells, and ~~on~~ of the long term fate of re-adsorbed SRP. Hence, all the SRP emitted from each cell through overland flow and sub-surface flow reaches the stream on the same day. For deep flow, only the immediate riparian flux is used in determining SRP inputs to the river.

No long-term depletion of the different P pools was modelled, because annual P export from the catchment was small compared to the size of soil and sub-soil P pools.

2.2.3 Input data and parameters

Spatial input data required for TNT2-P include:

- A DEM in raster format. Here, a 20 m resolution DEM was used, hence model calculations were made in 12348 grid cells covering a 4.94 km^2 catchment.
- A map of soil~~s~~ units that could be assumed to have~~with~~ homogeneous hydrological parameter values, in raster format. Here, two soil classes were considered by differentiating well-drained (86%) and ~~poorly~~ poorly-drained soils (14%) according to Curmi et al. (1998) (Fig. 1). Experimental determination of saturated hydraulic conductivity (29 soil cores) by Curmi et al. (1998) showed significantly different values for soils classified as well-drained and poorly-drained in the Kervidy-Naizin catchment. The two units were treated as homogeneous, lacking information about the detailed variability in soil hydraulic characteristics at the model grid scale.
- A map of surface Olsen P in raster format and description of decrease in ~~P Olsen~~Olsen P with depth for five soil layers between 0-50 cm. Here, the map of Olsen P in the 0-15 cm soil layer was obtained from statistical modelling with the rule-based regression algorithm CUBIST (Quinlan, 1992) using data from 198 soil samples (2013) in an area of 12 km^2 encompassing the 4.94 km^2 catchment (Matos-Moreira et al., 2015). To describe how ~~P Olsen~~Olsen P decreases with depth, land use information was used. In tilled fields, i.e. all crop rotations including arable crops, Olsen P was assumed to be constant between 0-30 cm and to decrease linearly with depth between 30-50 cm. In no-till fields, i.e. permanent pasture and woodland, Olsen P was assumed to decrease linearly with depth between 0-50 cm. An exponential decrease with depth is more commonly adopted in untilled land (e.g. Haygarth et al., 1998; Page et al., 2005), but a specific sampling in currently untilled areas in the Kervidy-Naizin catchment (Dupas et al., 2015c) has shown that a linear function is more appropriate, probably because of these areas having been ploughed in the past. A previous study has shown that soil Olsen P was the most important factor controlling SRP solubilisation in soils of the Kervidy-Naizin catchment (see section 2.1.3.), so other parameters in the soil-P sub-model (section 2.2.2.) were treated as homogeneous in the catchment (the soil classification into well-drained and poorly-drained soils only concerned hydrological parameters).

A 20 m resolution was chosen for the DEM and the soil Olsen P raster map to allow a detailed representation of the interaction of the groundwater table (as simulated by the hydrological model) and the soil Olsen P (as given by the soil Olsen P map). Indeed the soil saturation and soil Olsen P can be very different in a narrow zone close to the stream compared to upslope due to the presence of a 5 to 50 m unfertilized buffer zone with lower Olsen P compared to fertilized fields. The Olsen P value close to the stream has a determining influence on SRP transfer, because this area is the most frequently connected to the stream, so a coarser resolution of the raster maps would degrade representation of the system.

Climate input data include minimum and maximum air temperature, precipitation, potential evapotranspiration, global radiation on a daily basis. The TNT2 model allows for several climate zones to be considered, in which case a raster map of climate zone must be provided to the model. Here, only one climate zone is considered.

In total, the TNT2-P model includes 15 parameters for each soil type, i.e. 30 parameters in total if two soil drainage classes are considered. To reduce the number of model runs necessary to explore the parameter space using Monte Carlo simulations, several parameters were given fixed values, or a constant ratio between the two soil types was set (Table 1). In the hydrological sub-model, the parameters to vary were identified in a previous sensitivity analysis (Moreau et al., 2013). In the soil sub-model, all the parameters were varied.~~.~~

Finally, only 12 parameters were varied independently (see Table 1). Initial parameter ranges for the hydrological sub-model were based on ~~literature-derived~~ values from several previous studies in Western France (Moreau et al., 2013) and those for the soil sub-model were based on a preliminary manual trial and error procedure. The SRP concentration for deep flow water was based on actual measurement of SRP in the weathered schist (Dupas et al., 2015c). A constant flux value for domestic sources was set at the 1% percentile of the daily flux between 2007 and 2013 (Dupas et al., 2015b).

2.3 Deriving limits of acceptability from data uncertainty assessment

The Monte Carlo based Generalized Likelihood Uncertainty Estimation (GLUE) methodology has been widely used in hydrology and is described elsewhere (Beven and Freer, 2001a; Beven, 2006, 2009). Briefly, the rationale of GLUE is that many model structures and parameter sets can give "acceptable" results, according to one or several performance measures~~, due to equifinality~~. Hence, GLUE considers that all models that give acceptable results should be used for prediction. A key issue in GLUE is to decide on a performance threshold to define acceptable models; typically, modellers set a threshold value of a measure such as the Nash-Sutcliffe Efficiency based on their subjective appreciation of data uncertainty or on previously used values. To allow for a more explicit justification of the performance threshold values used, the limits of acceptability approach outlined by Beven (2006) relies on an assessment of uncertainty in the calibration/evaluation data. According to this approach, all model realisations that fall within the limits of acceptability are used for prediction, weighted by a score calculated based on overall performance.

Details on how the limits of acceptability for daily discharge and daily SRP load were derived from uncertainty assessment of the observational data are presented below. Input data, such as weather and soil Olsen P data, also contained uncertainties which were not accounted for explicitly in the limits of acceptability due to a lack of data to quantify them.

2.3.1 Discharge

Error in discharge measurement data was assessed from the original discharge measurements used to calibrate the stage-discharge rating curve (Carluer, 1998). The rating curve used in this study was:

$$Q = a * (h - h_0)^b \qquad (7)$$

Where Q is discharge, h is stage reading, h_0 is stage reading at zero discharge, a and b are calibrated coefficients. Limits of acceptability were defined as the 90% prediction interval of log-log linear regression (Fig. 3). The ~~Estimated~~ acceptability range estimated in this way was ±39% on average. This uncertainty interval is in the higher range of values found in other studies, e.g. Coxon et al. (2015) who found that mean discharge uncertainty was generally between 20% and 40% in 500 catchments of the United Kingdom. This relatively large uncertainty interval is due to the fact that it was derived from a prediction interval rather than a confidence interval (the 90% confidence interval of the log-log linear regression would be 14% of the mean discharge value during the study period). A prediction interval is an interval in which future observations will likely fall, while a confidence interval is an interval in which the mean of repeated observation will likely fall. Because in the TNT2-P model´s evaluation, we want each observation to fall in the acceptability interval (section 2.3.3.), a prediction interval was more appropriate. For daily discharge values below 2 mm d^{-1}, fixed acceptability limits were set at the 90% prediction interval for a stage measurement corresponding to 2 mm d^{-1}.

2.3.2 SRP load

Uncertainty in "observed" daily load includes uncertainty in discharge (see 2.3.1.) and uncertainty in SRP concentration. The acceptability limit for~~Uncertainty in~~ daily load was estimated summing up relative uncertainty assessed for discharge and SRP concentration (in percentage). Uncertainty in SRP concentration stems from sampling frequency problems as one grab sample collected on a specific day is incommensurable with the mean daily concentration or load simulated by the model. Further, measurement errors exist that include the effect of storage time (Haygarth et al., 1995). During baseflow periods, measurement error was expected to be the main source of uncertainty because relative measurement error is large for low concentrations, especially when sample storage time exceeds 48h (Jarvie et al., 2002), while concentrations vary little. During storm events, sampling frequency was expected to be the main source of uncertainty because SRP concentration can vary by one order of magnitude within a few hours. Therefore, different acceptability limits were set for both flow conditions. We considered storms as events with > 20 l s^{-1} increase in discharge and the following 24h.

During baseflow periods, the acceptability limits were derived from the 90% prediction interval of a linear regression model (y = a * x + b) linking pairs of data points sampled on the same day (reference sample between 16:00-18:00, verification sample between 11:00-15:00) and analysed independently (within a fortnight for the reference sample and within 1-2 days for the verification sample). It was assumed that there was no systematic bias between the two datasets due to different sampling time. The reference SRP concentrations were on average 13% lower than the verification value but this difference was not statistically significant (Mann-Whitney Rank Sum Test, $p > 0.05$). ~~Hence, the expected underestimation of SRP concentration due to long sample storage appears to be overshadowed by other sources of uncertainty such as variability in SRP concentration during the day of sampling or analytical imprecision at low concentrations.~~ This method encompasses all various sources of uncertainty, which results in prediction intervals much wider than what would result from a mere repeatability test: at the median concentration (0.02 mg l^{-1}), estimated prediction interval was 166% with this method versus 57% with a repeatability test (Fig. 4). As for discharge estimates, the high percentage represents a small absolute value (0.03 mg l^{-1}) during baseflow periods.

During storm events, acceptability limits were derived from the 90% prediction interval of concentration discharge ~~empirical~~ statistical models (C = a*Q^b) using high frequency autosampler data. Two reasons led us to use a statistical model (which also implies the assumption that errors are aleatory and temporally independent): i) the measurement uncertainty as assessed by the laboratory repetition test was an underestimate of the real uncertainty of autosampler data, because it does not include other major sources of error such as delayed filtration and sample decay during storage; ii) it was necessary to extrapolate the sub-daily observation to the daily resolution of the model. The limits of this choice will be discussed in section 4.3. An~~ distinct~~ empirical model was used to fit to each storm event monitored separately and a delay term was introduced manually in the empirical model when a time lag existed between concentration and discharge peaks. The empirical models were then applied to extrapolate concentration estimation during two days at 10 min resolution, for each of the 14 storm events monitored. Finally the 2-day mean "observed" load was estimated as the mean of 10 min loads and uncertainty limits were derived from the 90% prediction interval. In model evaluation, the mean of simulated loads during 2 consecutive days was evaluated against the 2-day mean "observed" load for which prediction intervals have been calculated. A 2-day acceptability limit enables ~~to cover the whole of~~all the storm events to be covered (Fig. 5 and Supplement). A 2-day aggregation was necessary here because increased SRP load as a response to each storm event could occur either mainly during the day of the rainfall (if the rainfall occurred early in the morning) or mainly during the day following the rainfall (if the rainfall occurred late in the evening), and with the daily resolution of the input data and model simulation, the information about the timing of the rainfall event was not available to the model.

When comparing autosampler data with data from immediately filtered samples, the ratio obtained had the range~~d~~ 1-1.6 (mean = 1.3), hence autosampler data were underestimates of the true concentration,~~d~~ arguably through adsorption or biological consumption. We used the mean ratio to correct all storm ~~uncertainty~~ acceptability intervals by 30% and the range values to extend the upper limit by 60%. During days with a storm event not monitored at high frequency with an autosampler, we considered that the grab sample data did not contain enough information to derive an acceptability interval for daily SRP load; hence simulated load was not evaluated for events not monitored at high frequency.~~.~~

2.3.3 Model runs and selection of acceptable models

To explore the parameter space, ~~15~~20,000 Monte Carlo realisations were performed to simulate daily discharge and SRP load during the water years 2013-2014 and 2014-2015. The number of Monte Carlo realisations was constrained by the computation time required to run a spatially explicit model in this catchment. A 7-month initialisation period was run to reduce the impact of initial conditions on simulated results during the study period, from 1 October 2013 to 31 July 2015.

To be considered acceptable, model runs must fall within the acceptability limits defined in 2.3.1 and 2.3.2. More specifically, 100% of simulated daily discharge, 100% of simulated baseflow SRP load and 100% of simulated storm SRP load had to fall within the acceptability limits. Thus, 572 acceptability tests were performed for discharge, 378 for baseflow SRP load and 14 for storm SRP loads, i.e. 964 evaluation criteria.

To evaluate the model performance in more detail, normalized scores were calculated during 6 periods (Table 2). To calculate the scores, a difference was calculated between each of the daily simulated discharge, baseflow SRP load and 2-day storm SRP loads and the corresponding observation. This difference was then normalized by the width of the acceptability limit defined for that day, so the score has a value of 0 in the case of a perfect match with observation, -1 at the lower limit and +1 at the upper limit (Fig. 6a). Finally, the median of this ratio was calculated for each of the 6 periods to investigate whether the model tended to underestimate or overestimate discharge and loads at different moments of the year and between the two years.

Model runs were successively evaluated for discharge, baseflow SRP load and storm SRP load. To use the models for prediction, each accepted model was given a likelihood weight according to how well it has performed for each of the 964 evaluation criteria. Here the statistical deviation weight was used (truncated to 90% prediction interval)~~a triangular weight was calculated for each evaluation criteria~~ (Fig. 5~~–~~b).~~, with the base of the triangle corresponding to the acceptability limit.~~ To "combine" the weights derived from the rating curve and the SRP concentration statistical models, a kernel density estimate (with Gaussian smoothing kernel) was computed to fit 10,000 realisations of the multiplied error models.

Calculated weights were then averaged for discharge, baseflow SRP load and storm SRP load respectively and the final likelihood was calculated as the ~~sum~~ product of all three averages.

The model's sensitivity to each hydrological and soil parameter was performed with a Hornberger-Spear-Young Generalised Sensitivity Analysis (HSY GSA, Whitehead and Young, 1979; Hornberger and Spear, 1981). For each evaluation criteria (daily discharge, daily baseflow SRP load, 2-day storm SRP load), the model runs were split into acceptable and non-acceptable runs according to the above-mentioned acceptability limits. Then a Kolmogorov-Smirnov test ~~is~~ was performed to assess whether the distribution of each of the three evaluation criteria differ between acceptable and non-acceptable models for each parameter. Because the Kolmogorov-Smirnov test might suggest that small differences in distribution are very significant when there are larger number of runs, this method is a qualitative guide to relative sensitivity. The p value of the Kolmogorov-Smirnov test is used to discriminate whether the model is critically sensitive ($p<0.01$ '***'), importantly sensitive ($p<0.1$ '*') or insignificantly sensitive ($p>0.1$ '.') to each parameter and for each of the three evaluation criteria. ~~Because the Kolmogorov-Smirnov test might suggest that small differences in distribution are very significant when there are larger number of runs, this method is a qualitative guide to relative sensitivity.~~

In addition to acceptability limit approach, a NSE (Moriasi et al., 2007) was calculated for daily discharge and daily load and concentration to allow comparison with other modelling studies where ~~is~~ it has been taken as an evaluation criterion.

3 Results

3.1 Presentation of observation data and calculation of acceptability limits

The two water years studied were highly contrasted in terms of hydrology and SRP loads. Water year 2013-2014 was the wettest in the last 10 years, with cumulative rainfall 1289 mm and cumulative runoff 716 mm. Water year 2014-2015 was an average year (5^{th} wettest in the last 10 years), with cumulative rainfall 677 mm and cumulative runoff 383 mm. Annual SRP load was 0.35 kg P ha^{-1} yr^{-1} in 2013-2014 and 0.17 kg P ha^{-1} yr^{-1} in 2014-2015, i.e. a difference 10% higher than that of discharge. Observed mean SRP concentration during the study period was 0.024 mg l^{-1}.

Fig. 7 a and b show~~s~~ acceptability limits for daily discharge and daily SRP loads. Note that acceptability limits for discharge were calculated every day, while acceptability limits for

SRP load was calculated on a daily basis during baseflow periods and on a 2-day basis during storm events monitored at high frequency. No SRP load acceptability limit was calculated during storm events when no high frequency autosampler data was available.

3.2 Model evaluation

First, model runs were evaluated against acceptability limits defined for discharge (Fig. 7c~~Fig. 8a~~). 5,479~~4,120~~/~~15~~20,000 models fulfilled the selection criterion for discharge, i.e. they had 100% of simulated daily discharge within the acceptability limits. The NSE estimated ~~for these~~for these models ranged from 0.~~78~~ 75 to 0.~~92~~93. The normalized scores calculated seasonally (Fig. 8~~9~~a) show that simulated discharge is often overestimated in autumn and spring, and underestimated in winter.

Then, model runs were evaluated against acceptability limits defined for SRP loads (Fig. 7d ~~Fig. 8b~~). During baseflow periods, 4,964~~3,730~~/20~~15~~,000 models fulfilled the selection criterion for SRP loads, i.e. they had 100% of simulated daily SRP load within the acceptability limits. Among them, 1,595~~1,210~~ also fulfilled the previous selection criterion for discharge. Normalized scores for baseflow SRP load showed the same trend as for discharge (Fig. ~~9b~~8b), i.e. overestimation in autumn and spring, and underestimation in winter. During storm events, only ~~5~~ 7 models fulfilled the selection criterion for SRP loads, i.e. they had 14/14 of simulated 2-day storm SRP loads within the acceptability limits, but none of them also fulfilled the selection criteria for discharge and baseflow SRP loads. Two storm events were particularly difficult to simulate (number 2 and number 9, Fig. ~~9c~~8c), probably because their acceptability interval was very narrow as a result of only small changes in discharge and concentration. To obtain a reasonable number of acceptable models, we relaxed the selection criterion so that the acceptable models had to simulate 12/14 of storm loads within the acceptability limits, in addition to the selection criteria defined for discharge and baseflow SRP load: ~~418~~ 539 models were then accepted. Estimated NSE of these ~~418~~ 539 models ranged from 0.09 to 0.~~80~~ 81 for daily load and from negative values to 0.53 for daily concentrations (this includes all data from the regular sampling).

3.3 Sensitivity analysis and prediction results

According to the HSA generalised sensitivity analysis, simulated discharge was critically sensitive to 10 out of the 12 hydrological parameters varied. Simulated SRP load was critically sensitive to the sub-surface and overland flow parameters during baseflow periods and to the overland flow parameter during storm events. During baseflow periods, SRP load was insignificantly sensitive to the parameter associated with deep flow load. Both baseflow and storm SRP loads were critically sensitive to the parameter related to soil moisture and soil temperature dependent SRP solubilisation (S1, T1 and T2), in addition to respectively ~~11~~ 12 and 8 hydrological parameters. This identification of sensitive parameters can be used in future application of the TNT2-P model in the study catchment, as suggested by Whitehead and Hornberger (1984) and Wade et al. (2002b).

Figure~~.~~ ~~10~~ 9 shows the daily discharge, SRP load and concentration as simulated by the acceptable models. Simulated SRP load during the water year 2013-2014 ranged 0.~~77~~ 81 – 3.25~~8~~ kg P ha^{-1} yr^{-1} (median = 1.68~~2~~ kg P ha^{-1} yr^{-1}); simulated SRP load during the water year 2014-2015 ranged 0.14 – 0.73 kg P ha^{-1} yr^{-1} (median = 0.34~~2~~ kg P ha^{-1} yr^{-1}). Best estimate of SRP load according to observation data was 0.35 kg P ha^{-1} yr^{-1} in 2013-2014 and 0.17 kg P ha^{-1} yr^{-1} in 2014-2015. According to the model, 49~~56~~ – 55~~61~~% (median = 52~~8~~%) of water discharge and 66~~71~~ – 70~~75~~% (median = 67~~2~~%) of SRP load occurred during storm events. Mean SRP concentrations during the two water years ranged 0.014~~3~~ – 0.044~~3~~ mg l^{-1} (median = 0.029~~8~~ mg l^{-1}), while mean observed SRP concentration was 0.024 mg l^{-1}.

4 Discussion

4.1 Role of hydrology and biogeochemistry in determining SRP transfer

The fairly good performance of TNT2-P at simulating SRP loads ~~confirms~~ provides further support that the hydrological and biogeochemical processes included into the model are dominant controlling factors in the Kervidy-Naizin catchment (i.e. the modelling hypotheses could not be rejected based on these results, expect for two storm events). The primary control of hydrology in controlling connectivity between soils and streams has been highlighted by many studies analysing water quality time series at the outlet of agricultural catchments (Haygarth et al., 2012; Jordan et al., 2012; Dupas et al., 2015c; Mellander et al., 2015). This modelling exercise also provides further support~~confirmed~~ that SRP solubility can be satisfactorily represented by~~was determined by~~ the soil ~~P Olsen~~Olsen P content and could vary according to temperature and moisture conditions. The underlying processes have not been identified precisely in the Kervidy-Naizin catchment: independent laboratory experiments have shown that microbial cell lysis resulting from alternating dry and water saturated periods in the soil could be the cause of increased SRP mobility (Turner and Haygarth, 2001; Blackwell et al., 2009). This could explain the moisture dependence of SRP solubility in the model. Furthermore, net mineralisation of soil organic phosphorus could explain the temperature dependence of SRP solubility in the model. These two hypotheses may explain increased SRP solubility in soils in periods of dry and hot conditions and will be further explored by incubation experiment with soils from the Kervidy-Naizin catchments.

4.2 Potential improvements to the model structure according to modelling purpose

The TNT2-P model was designed to test hypotheses about dominant processes and for this purpose, a parsimonious model structure was chosen to include only the processes which were to be tested. This parsimonious model structure might contain some conceptual misrepresentations due to oversimplification, and it might not include all the processes necessary for the purpose of evaluating management scenarios. This section discusses whether the simplifications made are acceptable in the context of different catchment types, and to which conditions the model could be made more complex by including additional routines for the purpose of evaluating management scenarios.

From a conceptual point of view, the lack of cell-to-cell routing of SRP fluxes might result in erroneous results in some contexts. The fact that all the SRP emitted from each cell through overland flow and sub-surface flow reaches the stream on the same day is generally acceptable for the catchment studied because groundwater interception of shallow soil layers occurs in the riparian zone only, hence the signal of SRP mobilisation in these soils is generally transmitted to the stream (Dupas et al., 2015c). This simplification, however, does not seem to be acceptable for all the storm events in the study catchment, as the SRP load evaluation criteria had to be relaxed to obtain acceptable model results. It would also not be acceptable in catchments where soil-groundwater interactions are taking place throughout the landscape, e.g. due to topographic depressions or poorly drained soils. In the latter type of catchment, transmission of the SRP mobilisation signal to the stream is more complex ~~to comprehend~~ (Haygarth et al., 2012)~~,~~; hence a more complex model structure would be required.

The reason for this simplification was that we lacked knowledge of SRP re-adsorption in downslope cells (or on suspended sediments in the stream network) and on the long-term fate of re-adsorbed SRP. For a more physically realistic representation of processes, it is likely that an explicit representation of flow velocities and pathways would be necessary, along with an explicit representation of several soil P pools. However, such an explicit representation of processes contradicts the idea of a parsimonious model, which was adopted here for the purpose of identifying dominant processes. In this respect, TNT2-P is an aggregative model rather than a fully distributed model although it is based on a fully distributed hydrological model (Beaujouan et al., 2002). The current spatial distribution allows finer representation of soil-groundwater interactions (i.e. the time varying extent of the riparian wetland area) than semi-distributed models such as SWAT (Arnold et al., 1998), INCA-P (Wade et al., 2002) and HYPE (Lindstrom et al., 2010) but at higher computational cost. It would be interesting to test to what~~ich~~ extent moving from an aggregative model with fully distributed information to a semi-distributed model would degrade the model performance while~~and in the same time~~ reducing~~e~~ computational cost.~~–~~ This could be achieved by grouping cells according to a hydrological similarity criterion like in ~~the original TOPMODEL and~~ Dynamic Topmodel (Beven and Freer, 2001b; Metcalfe et al., 2015) and do the same for similarity in soil P content. Reducing computation time is critical in the context of a GLUE analysis because this method requires the parameter space to be sampled adequately to identify those models to be considered acceptable. This is debatable here because 12 parameters were varied and only 20,000 model runs were performed. It is therefore possible that some regions of the parameter space with acceptable models might not have been sampled.

If reducing the number of calculation units proved to reduce computational cost without degrading quality of prediction, it would be possible to include more parameters in the model, for example to simulate SRP re-absorption in downslope cells or include routines to simulate the evolution of soil P content under different management scenarios (Vadas et al., 2011; 2012), and still perform a Monte-Carlo based analysis of uncertainty. The question of coupling or not such a soil P routine with the current TNT2-P model will depend on available data and on the length of available time series: studying the evolution of the soil P content requires at least a decade of soil observation data (Ringeval et al., 2014) and probably a longer period of stream data to account for the time delay for a perturbation in the catchment to become visible in the stream (Wall et al., 2013). Thus, the two years of daily stream SRP in the Kervidy-Naizin catchment are not enough to build a coupled soil-hydrology model with an elaborate soil P routine. Therefore, as things stand, it is more reasonable to generate new soil ~~P Olsen~~Olsen P maps with a separate model such as the APLE model (Vadas et al., 2012;

Benskin et al., 2014) or the 'soil P decline' model used by Wall et al. (2013), and use these maps as input to TNT2-P.

Because the current model can simulate response to rainfall, soil moisture and temperature, it could be used to test the effect of climate scenarios on SRP transfer. In Western France, and more generally in Western Europe, the climate for the next few decades is expected to consist of hotter, drier summers and warmer, wetter winter (Jacob et al., 2007; Macleod et al., 2012; Salmon-Monviola et al., 2013) with increased frequency of high intensity rainfall events (Dequé 2007). In these conditions, SRP concentrations and load will seemingly increase compared to today's climate as a result of both an increase in SRP solubility in soil due to higher temperature and more severe drought and an increase in transfer due to wetter winter and more frequent high intensity rainfall events. TNT2-P could be used to confirm and quantify the expected increase in SRP transfer from diffuse sources in future climate scenarios, and to determine whether those predicted changes are significant relative to the uncertainty in predictions under current climate variability.~~conditions.~~

4.3 Improving information content in the data

Despite relatively large uncertainty in the data used in this study, it was possible to build a parsimonious catchment model of SRP transfer for the purpose of testing hypotheses about dominant processes, namely the role of hydrology in controlling connectivity between soils and streams and the role of temperature and moisture conditions in controlling soil SRP solubilisation. However, the large uncertainties in the calibration data lead to large prediction uncertainty. For example, the SRP load estimated by the behavioural models from 2013 to 2015 ranged from 0.48~~5~~ to 1.99~~2.0~~ kg P ha^{-1} yr^{-1}; hence the width of the credibility interval was 15~~6~~0% of the median (1.00~~0.97~~ kg P ha^{-1} yr^{-1}). Similarly, the mean SRP concentration estimated by the behavioural models from 2013 to 2015 ranged from 0.01~~3~~4 to 0.04~~4~~5 mg l^{-1}; hence the width of the credibility interval was 102~~10~~% of the median (0.02~~8~~9 mg l^{-1}). The large uncertainty in the calibration data, along with a lack of long-term information, also prevents including more detailed processes in the soil routine.

To reduce uncertainty in prediction and to build more complex models, several options exist to improve information content in the data. As stated by Jackson-Blake et al. (2015b), "the key to obtaining a realistic model simulation is ensuring that the natural variability in water chemistry is well represented by the monitoring data". The monitoring strategy adopted in the

Kervidy-Naizin catchment should theoretically enable to capture the natural variability in stream SRP concentration, because sampling took place during two contrasting water years, during different seasons and at a high frequency during 14 storm events. The analysis of uncertainty in the data shows that a large part of uncertainty in "observed" SRP concentration originates from sample storage, both unfiltered between the time of autosampling and manual filtration and between filtration and analysis. This is due to SRP being non-conservative. Thus, there is room for improvement in reducing storage time, without increasing further the monitoring frequency. In this respect, the primary interest of investing in high frequency bankside analysers would lie in their ability to analyse water samples immediately in addition to providing near continuous data. Because bankside analysers perform measurements in relatively homogeneous conditions, unlike the manual and autosampler data for which storage time of filtered and unfiltered samples vary, a finer quantification of uncertainty in the measurement data would be possible (e.g. Lloyd et al., ~~2015~~2016).

Finally, alternative methods to statistical models could be used to derive acceptability limits (in this study three statistical models are used: the rating curve, the SRP concentration uncertainty during baseflow periods and the storm event interpolation model) because statistical models have at least three shortcomings: i) they lump the uncertainty linked to the timing of sampling, the immediate or delayed filtration of the samples, the storage time and the analytical error; ii) the formula chosen adds error to the already existing measurement errors because empirical models are not perfect representation of the system dynamics; iii) they assume a parametric distribution and temporally independent errors which are not always verified in practice. As an alternative, non-parametric methods could be used, but these methods generally require a large number of data points and they are not suitable for extrapolation to extreme values.

5 Conclusion

The TNT2-P model was capable of capturing daily variation of SRP loads, thus confirming the dominant processes identified in previous analyses of observation data in the Kervidy-Naizin catchment. The role of hydrology in controlling connectivity between soils and streams, and the role of soil Olsen P, soil moisture and temperature in controlling SRP solubility have been confirmed. The lack of any representation of the short-term effect of management practices did not seem to penalize the model's performance. Their long-term effect on the soil Olsen P could be simulated with an independent model or through an additional sub-model if a longer period of data was available to calibrate it. The modelling approach presented in this paper included an assessment of the information content in the data, and propagation of uncertainty in the model's prediction. The information content of the data was sufficient to explore dominant processes, but the relatively large uncertainty in SRP concentrations would seemingly limit the possibility for including more detailed processes into the model. Data from near continuous bankside analyser will probably allow calibrating more detailed models in the near future.

References

Alexander RB, Smith RA, Schwarz GE, Boyer EW, Nolan JV, Brakebill JW. Differences in phosphorus and nitrogen delivery to the gulf of Mexico from the Mississippi river basin. Environmental Science & Technology 2008; 42: 822-830.

Arnold JG, Srinivasan R, Muttiah RS, Williams JR. Large area hydrologic modeling and assessment - Part 1: Model development. Journal of the American Water Resources Association 1998; 34: 73-89.

Aubert AH, Gascuel-Odoux C, Gruau G, Akkal N, Faucheux M, Fauvel Y, et al. Solute transport dynamics in small, shallow groundwater-dominated agricultural catchments: insights from a high-frequency, multisolute 10 yr-long monitoring study. Hydrology and Earth System Sciences 2013; 17: 1379-1391.

Beauchemin S, Simard RR. Soil phosphorus saturation degree: Review of some indices and their suitability for P management in Quebec, Canada. Canadian Journal of Soil Science 1999; 79: 615-625.

Beaujouan V, Durand P, Ruiz L, Aurousseau P, Cotteret G. A hydrological model dedicated to topography-based simulation of nitrogen transfer and transformation: rationale and application to the geomorphology-denitrification relationship. Hydrological Processes 2002; 16: 493-507.

Benskin CMH, Roberts W. M, Wang Y, Haygharth PM. Review of the Annual Phosphorus Loss Estimator tool – a new model for estimating phosphorus losses at the field scale. Soil Use and Management 2014; 30: 337-341.

Beven K. A manifesto for the equifinality thesis. Journal of Hydrology 2006; 320: 18-36.

Beven K. Environmental Modelling – An Uncertain Future? Routledge: London 2009.

Beven K, Freer J. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 2001a; 249: 11-29.

Beven K, Freer J. A dynamic TOPMODEL. Hydrological Processes 2001b; 15: 1993-2011.

Beven K, Smith P. Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models. Journal of Hydrologic Engineering 2015; 20.

Beven KJ. Distributed hydrological modelling: applications of the TOPMODEL concept, 1997.

Blackwell MSA, Brookes PC, de la Fuente-Martinez N, Murray PJ, Snars KE, Williams JK, et al. Effects of soil drying and rate of re-wetting on concentrations and forms of phosphorus in leachate. Biology and Fertility of Soils 2009; 45: 635-643.

Blazkova S, Beven K. A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. Water Resources Research 2009; 45.

~~Bradley WG, Borenstein AR, Nelson LM, Codd GA, Rosen BH, Stommel EW, et al. Is exposure to cyanobacteria an environmental risk factor for amyotrophic lateral sclerosis and other neurodegenerative diseases? Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 2013; 14: 325-333.~~

Bruneau P, Gascuel-Odoux C, Robin P, Merot P, Beven KJ. Sensitivity to space and time resolution of a hydrological model using digital elevation data. Hydrological Processes 1995; 9: 69-82.

Carluer N. Vers une modélisation hydrologique adaptée à l'évaluation des pollutions diffuses: prise en compte du réseau anthropique. Application au bassin versant de Naizin (Morbihan). PhD thesis Université Pierre et Marie Curie 1998.

Carpenter SR, Caraco NF, Correll DL, Howarth RW, Sharpley AN, Smith VH. Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecological Applications 1998; 8: 559-568.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531-5546, 2015.

Curmi P, Durand P, Gascuel-Odoux C, Merot P, Walter C, Taha A. Hydromorphic soils, hydrology and water quality: spatial distribution and functional modelling at different scales. Nutrient Cycling in Agroecosystems 1998; 50: 127-142.

Dean S, Freer J, Beven K, Wade AJ, Butterfield D. Uncertainty assessment of a process-based integrated catchment model of phosphorus. Stochastic Environmental Research and Risk Assessment 2009; 23: 991-1010.

Deque M. Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. Global and Planetary Change 2007; 57: 16-26.

Dupas R, Delmas M, Dorioz JM, Garnier J, Moatar F, Gascuel-Odoux C. Assessing the impact of agricultural pressures on N and P loads and eutrophication risk. Ecological Indicators 2015a; 48: 396–407.

Dupas R, Gascuel-Odoux C, Gilliet N, Grimaldi C, Gruau G. Distinct export dynamics for dissolved and particulate phosphorus reveal independent transport mechanisms in an arable headwater catchment. Hydrological Processes 2015b.

Dupas R, Gruau G, Gu S, Humbert G, Jaffrezic A, Gascuel-Odoux C. Groundwater control of biogeochemical processes causing phosphorus release from riparian wetlands. Water Research 2015c.

Dupas R, Tavenard R, Fovet O, Gilliet N, Grimaldi C, Gascuel-Odoux C. Identifying seasonal patterns of phosphorus storm dynamics with Dynamic Time Warping. Water Resources Research 2015d.

Durand P, Moreau P, Salmon-Monviola J, Ruiz L, Vertes F, Gascuel-Odoux C. Modelling the interplay between nitrogen cycling processes and mitigation options in farming catchments. Journal of Agricultural Science 2015; 153: 959-974.

Franks SW, Gineste P, Beven KJ, Merot P. On constraining the predictions of a distributed model: the incorporation of fuzzy estimates of saturated areas into the calibration process, Water Resources Research 1998; 34: 787-797.

Grizzetti B, Bouraoui F, Aloe A. Changes of nitrogen and phosphorus loads to European seas. Global Change Biology 2012; 18: 769-782.

Hahn C, Prasuhn V, Stamm C, Lazzarotto P, Evangelou MWH, Schulin R. Prediction of dissolved reactive phosphorus losses from small agricultural catchments: calibration and validation of a parsimonious model. Hydrology and Earth System Sciences 2013; 17: 3679-3693.

Hahn C, Prasuhn V, Stamm C, Schulin R. Phosphorus losses in runoff from manured grassland of different soil P status at two rainfall intensities. Agriculture Ecosystems & Environment 2012; 153: 65-74.

Haygarth PM, Ashby CD, Jarvis SC. Short-term changes in the molybdate reactive phosphorus of stored soil waters. Journal of Environmental Quality 1995; 24: 1133-1140.

Haygarth PM, Hepworth L, Jarvis SC. Forms of phosphorus transfer in hydrological pathways from soil under grazed grassland. European Journal of Soil Science 1998; 49: 65-72.

Haygarth PM, Page TJC, Beven KJ, Freer J, Joynes A, Butler P, et al. Scaling up the phosphorus signal from soil hillslopes to headwater catchments. Freshwater Biology 2012; 57: 7-25.

Heathwaite AL, Dils RM. Characterising phosphorus loss in surface and subsurface hydrological pathways. Science of the Total Environment 2000; 251: 523-538.

Heckrath G, Brookes PC, Poulton PR, Goulding KWT. Phosphorus leaching from soils containing different phosphorus concentrations in the broadbalk experiment. Journal of Environmental Quality 1995; 24: 904-910.

Hornberger GM, Spear RC. An approach to the preliminary analysis of environmental systems. J. Environmental Management 1981; 12: 7-18.

Jackson-Blake LA, Dunn SM, Helliwell RC, Skeffington RA, Stutter MI, Wade AJ. How well can we model stream phosphorus concentrations in agricultural catchments? Environmental Modelling & Software 2015a; 64: 31-46.

Jackson-Blake LA, Starrfelt J. Do higher data frequency and Bayesian auto-calibration lead to better model calibration? Insights from an application of INCA-P, a process-based river phosphorus model. Journal of Hydrology 2015b; 527: 641-655.

Jacob D, Barring L, Christensen OB, Christensen JH, de Castro M, Deque M, et al. An inter-comparison of regional climate models for Europe: model performance in present-day climate. Climatic Change 2007; 81: 31-52.

Jarvie HP, Withers PJA, Neal C. Review of robust measurement of phosphorus in river water: sampling, storage, fractionation and sensitivity. Hydrology and Earth System Sciences 2002; 6: 113-131.

Jordan P, Melland AR, Mellander PE, Shortle G, Wall D. The seasonality of phosphorus transfers from land to water: implications for trophic impacts and policy evaluation. Sci Total Environ 2012; 434: 101-9.

Kirchner JW. Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research 2006; 42.

Krueger T, Quinton JN, Freer J, Macleod CJA, Bilotta GS, Brazier RE, Hawkins JMB, Haygarth PM. Comparing empirical models for sediment and phosphorus transfer from soils to water at field and catchment scale under data uncertainty. European Journal of Soil Science 2012; 63(2): 211–223.

Humbert G, Jaffrezic A, Fovet O, Gruau G, Durand P. Dry-season length and runoff control annual variability in stream DOC dynamics in a small, shallow groundwater-dominated agricultural watershed. Water Resources Research 2015.

Lazzarotto P, Stamm C, Prasuhn V, Flühler H. A parsimonious soil-type based rainfall-runoff model simultaneously tested in four small agricultural catchments. Journal of Hydrology 2006; 321: 21-38.

Lindstrom G, Pers C, Rosberg J, Stromqvist J, Arheimer B. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. Hydrology Research 2010; 41: 295-319.

Lloyd CEM, Freer JE, Johnes PJ, Coxon G, Collins AL. Discharge and nutrient uncertainty: implications for nutrient flux estimation in small streams. Hydrological processes ~~2015~~2016; 30: 165-152.

Macleod CJA, Falloon PD, Evans R, Haygarth PM. The effects of climate change on the mobilization of diffuse substances from agricultural systems. In: Sparks DL, editor. Advances in Agronomy, Vol 115. 115, 2012, pp. 41-77.

Maguire RO, Sims JT. Soil testing to predict phosphorus leaching. Journal of Environmental Quality 2002; 31: 1601-1609.

Matos-Moreira M, Lemercier B, Michot D, Dupas R, Gascuel-Odoux C. Using agricultural practices information for multiscale environmental assessment of phosphorus risk. Geophysical Research Abstracts 2015; 17.

McDowell R, Sharpley A, Withers P. Indicator to predict the movement of phosphorus from soil to subsurface flow. Environmental Science & Technology 2002; 36: 1505-1509.

McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality, Hydrological Processes, 26, 4078-4111, 2012.

Mellander PE, Jordan P, Shore M, Melland AR, Shortle G. Flow paths and phosphorus transfer pathways in two agricultural streams with contrasting flow controls. Hydrological Processes 2015.

Metcalfe P, Beven BJ, and Freer J. Dynamic Topmodel: a new implementation in R and its sensitivity to time and space steps. Environmental Modelling and Software 2015; 72: 155-172.

Molenat J, Gascuel-Odoux C, Ruiz L, Gruau G. Role of water table dynamics on stream nitrate export and concentration. in agricultural headwater catchment (France). Journal of Hydrology 2008; 348: 363-378.

Moore MT, Locke MA. Effect of Storage Method and Associated Holding Time on Nitrogen and Phosphorus Concentrations in Surface Water Samples. Bulletin of Environmental Contamination and Toxicology 2013; 91: 493-498.

Moreau P, Ruiz L, Mabon F, Raimbault T, Durand P, Delaby L, et al. Reconciling technical, economic and environmental efficiency of farming systems in vulnerable areas. Agriculture Ecosystems & Environment 2012; 147: 89-99.

Moreau P, Viaud V, Parnaudeau V, Salmon-Monviola J, Durand P. An approach for global sensitivity analysis of a complex environmental model to spatial inputs and parameters: A case study of an agro-hydrological model. Environmental Modelling & Software 2013; 47: 74-87.

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the Asabe 2007; 50: 885-900.

Olsen SR, Cole CV, Watanbe FS, Dean LA. Estimation of available phosphorus in soils by extraction with sodium bicarbonate 1954.. Circ. 939. USDA, Washington, DC.

Outram FN, Lloyd CEM, Jonczyk J, Benskin CMH, Grant F, Perks MT, et al. High-frequency monitoring of nitrogen and phosphorus response in three rural catchments to the end of the 2011-2012 drought in England. Hydrology and Earth System Sciences 2014; 18: 3429-3448.

Page T, Haygarth PM, Beven KJ, Joynes A, Butler T, Keeler C, et al. Spatial variability of soil phosphorus in relation to the topographic index and critical source areas: Sampling for assessing risk to water quality. Journal of Environmental Quality 2005; 34: 2263-2277.

Perks MT, Owen GJ, Benskin CMH, Jonczyk J, Deasy C, Burke S, et al. Dominant mechanisms for the delivery of fine sediment and phosphorus to fluvial networks draining grassland dominated headwater catchments. Science of the Total Environment 2015; 523: 178-190.

Quinlan, J.R. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference On Artificial Intelligence 1992, 343-348.

Rode M, Suhr U. Uncertainties in selected river water quality data. Hydrology and Earth System Sciences 2007; 11(2): 863–874.

Ringeval B, Nowak B, Nesme T, Delmas M, Pellerin S. Contribution of anthropogenic phosphorus to agricultural soil fertility and food production. Global Biogeochemical Cycles 2014; 28: 743-756.

Salmon-Monviola J, Moreau P, Benhamou C, Durand P, Merot P, Oehler F, et al. Effect of climate change and increased atmospheric CO2 on hydrological and nitrogen cycling in an intensive agricultural headwater catchment in western France. Climatic Change 2013; 120: 433-447.

Schindler DW, Hecky RE, Findlay DL, Stainton MP, Parker BR, Paterson MJ, et al. Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. Proceedings of the National Academy of Sciences of the United States of America 2008; 105: 11254-11258.

Schoumans OF, Chardon WJ. Phosphate saturation degree and accumulation of phosphate in various soil types in The Netherlands. Geoderma 2015; 237: 325-335.

Serrano T, Dupas R, Upegui E, Buscail C, Grimaldi C, Viel J-F. Geographical modeling of exposure risk to cyanobacteria for epidemiological purposes. Environment International 2015; 81: 18-25.

Sharpley AN, Kleinman PJ, Heathwaite AL, Gburek WJ, Folmar GJ, Schmidt JP. Phosphorus loss from an agricultural watershed as a function of storm size. J Environ Qual 2008; 37: 362-8.

Siwek J, Siwek JP, Zelazny M. Environmental and land use factors affecting phosphate hysteresis patterns of stream water during flood events (Carpathian Foothills, Poland). Hydrological Processes 2013; 27: 3674-3684.

Turner BL, Haygarth PM. Biogeochemistry - Phosphorus solubilization in rewetted soils. Nature 2001; 411: 258-258.

Vadas PA, Joern BC, Moore PA. Simulating soil phosphorus dynamics for a phosphorus loss quantification tool. J Environ Qual 2012; 41: 1750-7.

Vadas PA, Jokela WE, Franklin DH, Endale DM. The Effect of Rain and Runoff When Assessing Timing of Manure Application and Dissolved Phosphorus Loss in Runoff1. JAWRA Journal of the American Water Resources Association 2011; 47: 877-886.

Wade AJ, Whitehead PG, Butterfield D. The Integrated Catchments model of Phosphorus dynamics (INCA-P), a new approach for multiple source assessment in heterogeneous river systems: model structure and equations. Hydrology and Earth System Sciences 2002; 6: 583-606.

Wall DP, Jordan P, Melland AR, Mellander PE, Mechan S, Shortle G. Forecasting the decline of excess soil phosphorus in agricultural catchments. Soil Use and Management 2013; 29: 147-154.

Whitehead PG, Hornberger GE. Modelling algal behaviour in the River Thames, Water Research 1984: 18: 945-953.

Wade AJ, Whitehead PG, Hornberger GE, Snook D. On Modelling the flow controls on macrophytes and epiphyte dynamics in a lowland permeable catchment: the River Kennet, southern England. Sci Tot Environ 2002b: 282-283: 395-417.

Whitehead P, Young P. Water-quality in river systems – Monte-Carlo analysis. Water Resources Research 1979; 15: 451-459.

Acknowledgements

This work was funded by the ‘‘Agence de l’Eau Loire Bretagne’’ via the ‘‘Trans-P project’’. Long-term monitoring in the Kervidy-Naizin catchment is supported by ‘‘ORE AgrHyS’’. Data of ‘‘ORE AgrHyS’’ can be downloaded from http://www6.inra.fr/ore_agrhys/Donnees.

Table 1: Initial parameter ranges in the hydrological and soil phosphorus sub models.

	Abbreviation	Unit	Hydrological (H), Phosphorus model (P)	Range poorly drained soils (min-max)	Range well drained soils (min-max)
Lateral transmissivity at saturation	T	$m^2\ d^{-1}$	H	4-8	-> x1.5
Exponential decay rate of hydraulic conductivity with depth	m	$m^2\ d^{-1}$	H	0.02-0.2	0.02-0.2
Soil depth	ho	m	H	0.3-0.8	-> x1
Drainage porosity of soil	po	$cm^3\ cm^{-3}$	H	0.1-0.4	-> x1
Regolith layer thickness	h1	m	H	5-10	-> x4
Exponent for evaporation limit	A	-	H	8 (fixed)	-> x1
kRC parameter for capillary rise	kRC	-	H	0.001 (fixed)	-> x1
n parameter for capillarity rise	N	-	H	2.5 (fixed)	-> x1
Drainage porosity of regolith layer	p1	$cm^3\ cm^{-3}$	H	0.01-0.05	-> x1
Background P release coefficient for subsurface flow	$Coef_{SRP\ overland}$	-	P	0-0.015	-> x1
Background P release coefficient for overland flow	$Coef_{SRP\ sub\text{-}surface}$	-	P	0-0.25	-> x1
Temperature coefficient 1	T1	-	P	5-10	-> x1
Temperature coefficient 2	T2	-	P	2-10	-> x1

Soil moisture coefficient	S1	-	P	0-2	-> x1
SRP concentration in deep flow	SRP_deep	mg l^{-1}	P	0-0.007	-> x1

Table 2: Starting and ending dates of periods studied

Name	Starting date	Ending date
Autumn 2013	01 October 2013	31 December 2013
Winter 2014	01 January 2014	31 March 2014
Spring 2014	01 April 2014	31 July 2014
Autumn 2014	01 October 2014	31 December 2014
Winter 2015	01 January 2015	31 March 2015
Spring 2015	01 April 2015	31 July 2015

Table 3: Sensitivity analysis of the model to 18 model parameters (insignificant ., important *, critical ***). Parameters significations are detailed in Table 1.

	discharge	**baseflow SRP load**	**storm SRP load**
T (poorly drained soils)	.	***	***
m (poorly drained soils)	***	***	***
ho (poorly drained soils)	***	***	.
po (poorly drained soils)	***	***	***
h1 (poorly drained soils)	***	***	.
p1 (poorly drained soils)	***	***	***
T (well drained soils)	.	***	***
m (well drained soils)	***	***	***
ho (well drained soils)	***	***	.
po (well drained soils)	***	***	***
h1 (well drained soils)	***	***	.
p1 (well drained soils)	***	***	***
Coef_sub-surface	.	***	.
Coef_overland	.	***	***
SRP_deep	.	.	.
S1	.	***	***
T1	.	***	***
T2	.	***	***

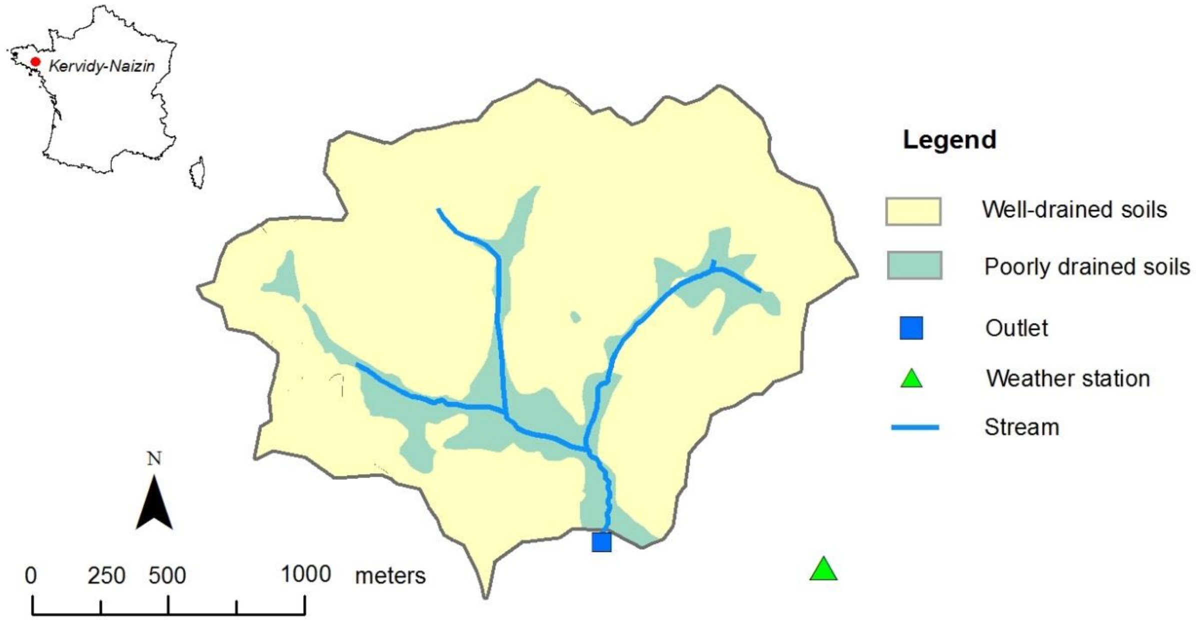

Fig. 1. Soil drainage classes in the Kervidy-Naizin catchment, Curmi et al. (1998)

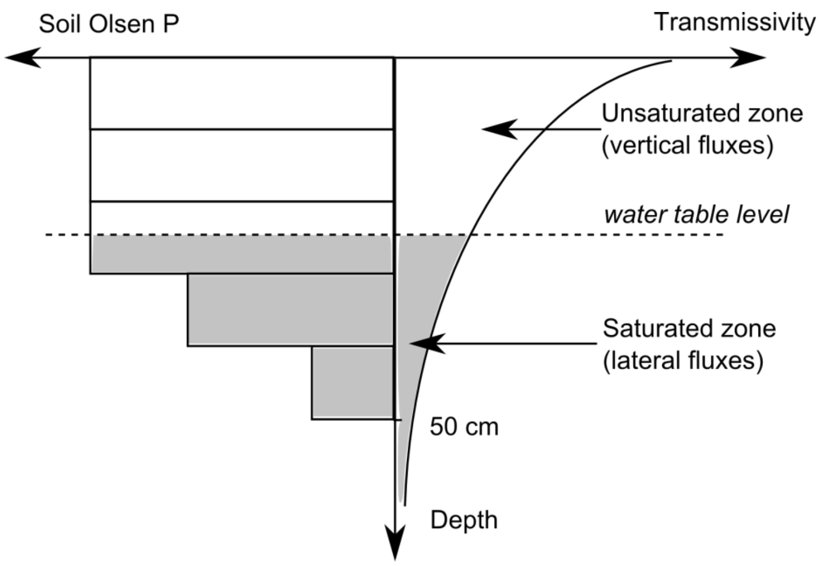

Fig. 2. Description of soil hydraulic properties and phosphorus content with depth

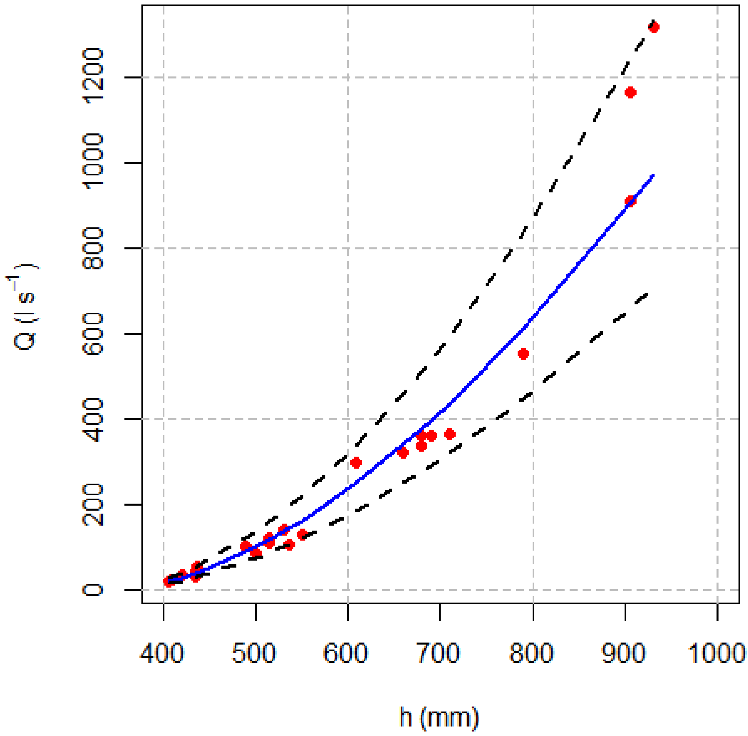

Fig. 3 : Rating curve in Kervidy-Naizin; acceptability bounds derived from 90% prediction interval (blue line: fitting regression; black dots: 90% prediction interval). Red dots represent the original discharge measurements used to calibrate the stage-discharge rating curve (Carluer, 1998).

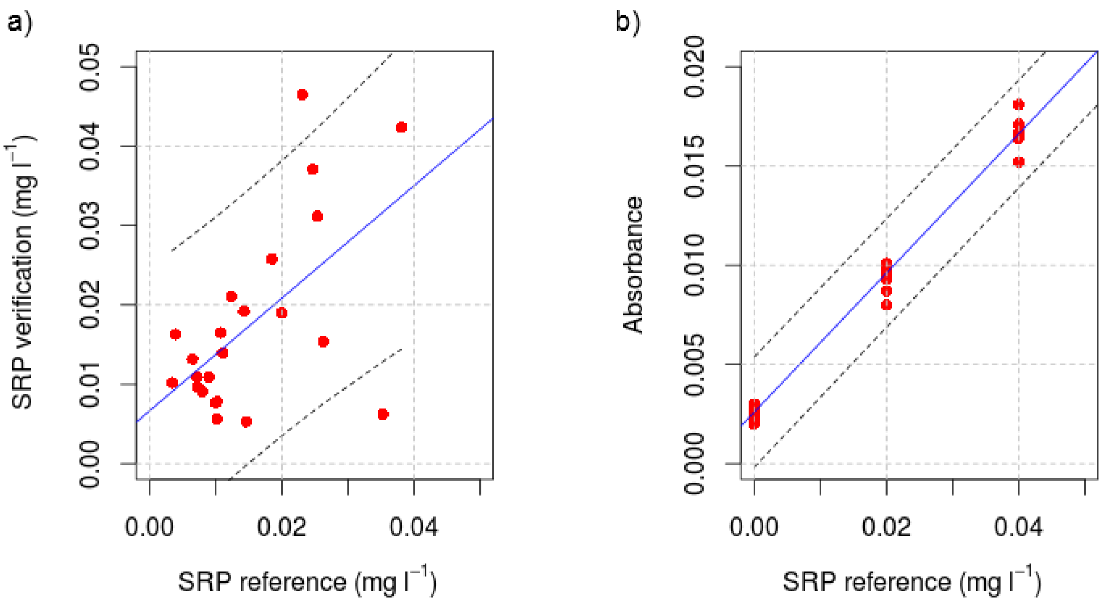

Fig. 4: a) linear regression model linking the reference data and a verification dataset; b) measurement error as estimated from a repeatability test performed by the lab in charge of producing reference data (blue line: fitting regression; black dots: 90% prediction interval).

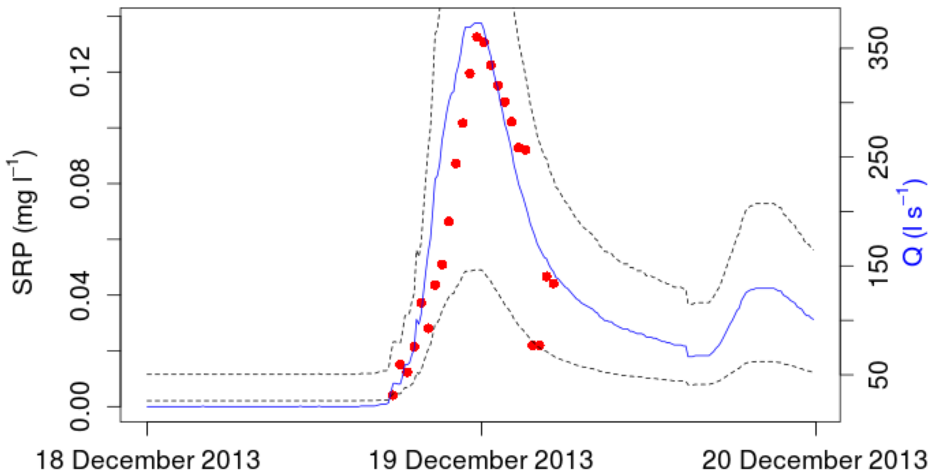

Fig. 5: Example of an empirical concentration – discharge model; acceptability bounds derived from 90% prediction interval. Red circles represent the SRP measurements.

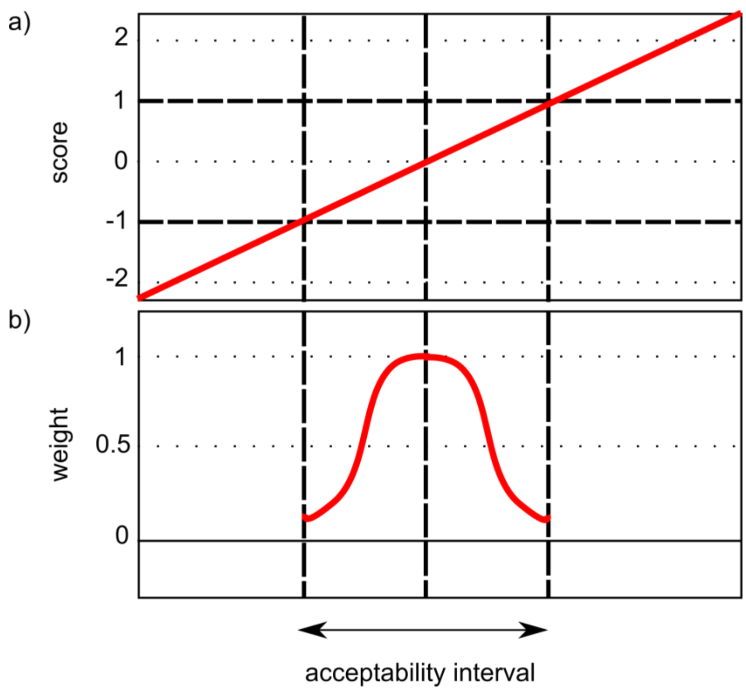

Fig. 6 : a) normalized scores; b) ~~triangular~~ weighting function

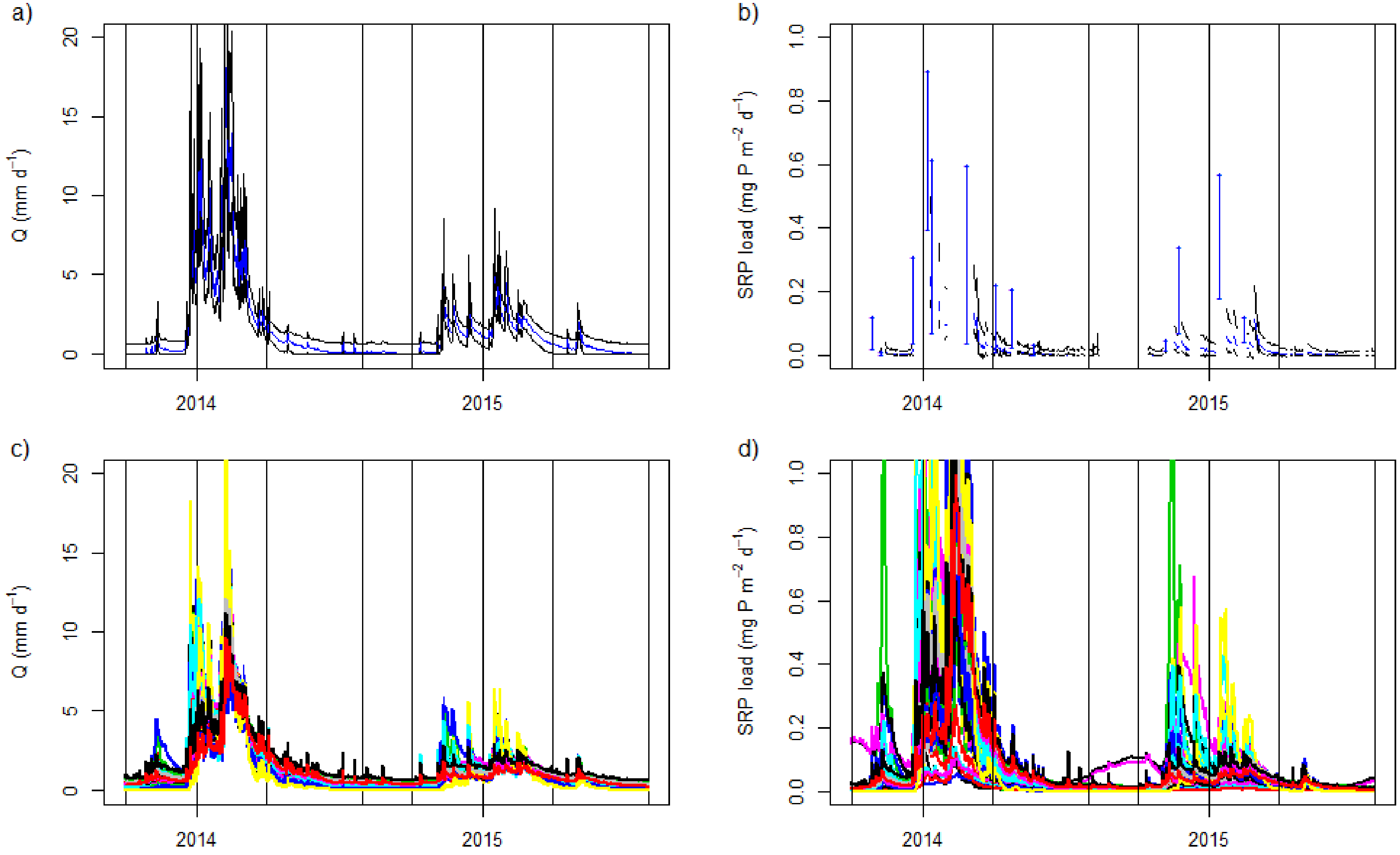
a)
b)
c)
d)
Q (mm d-1)
SRP load (mg P m-2 d-1)
2014
2015

Fig. 7: Acceptability limits for daily discharge (a) and SRP load (b). Blue lines represent best estimates; black lines represent the acceptability limits. Storm loads acceptability limits are represented by vertical blue lines. And example of 50 model runs simulating discharge (c) and daily load (d). Black vertical lines represent the starting and ending dates for each season (table 2).

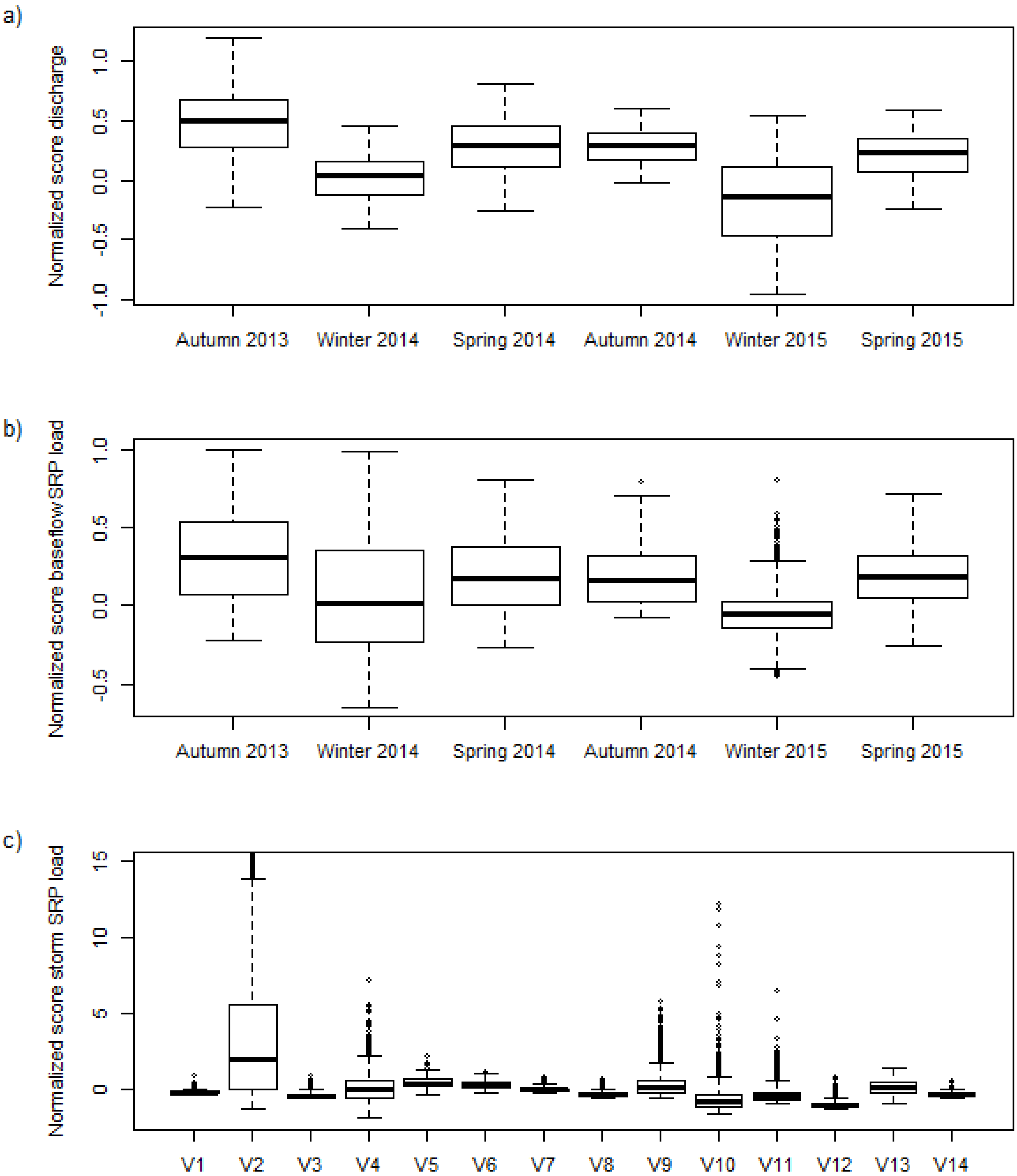

Fig. ~~9~~8: Normalized score for daily discharge (a), baseflow SRP load (b) and storm SRP load (c).

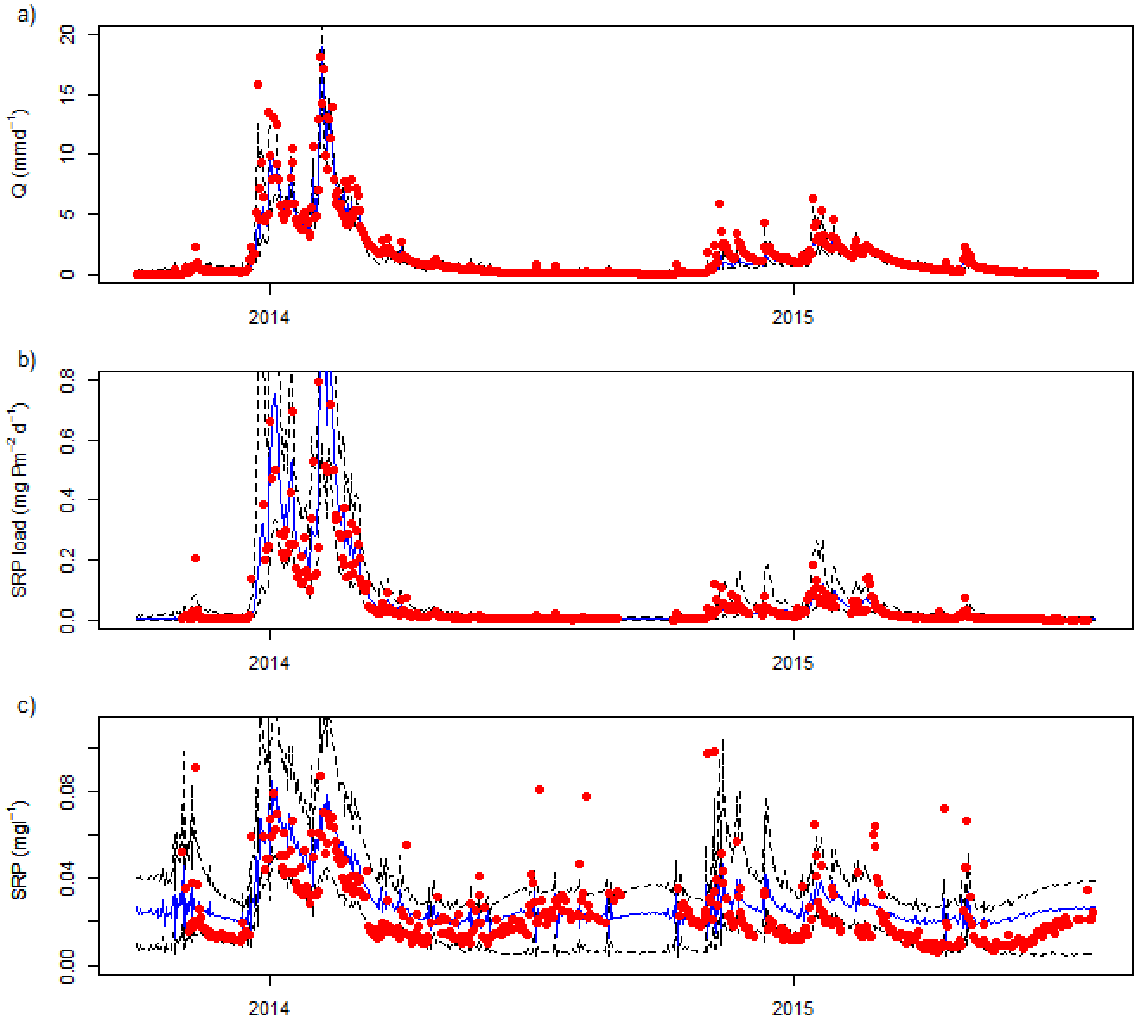

Fig. ~~10~~9: Median and 95% credibility interval for daily discharge (a), SRP load (b) and SRP concentration (c). Red circles represent observational data.