Authors' response to Reviewer 1, Sebastian Stoll
For clarity, we have included the reviewer's comments in black; our response is in blue

General remarks:
Generally, I find the manuscript to be very interesting, well written and suitable for HESS (after some revisions).

Thank you

While I agree with the authors that DBM models are very helpful in detecting dominant transfer modes I think that some of the alleged benefits of the modelling approach are overstated. For example, I doubt that these models can "help in planning appropriate pollution mitigation measures" as stated in the abstract. The reason for that is the nature of these models. The only input driving the models is rainfall (and sometimes discharge) data. Many features which are known to influence the phosphorus dynamics (like soil type, soil phosphorus concentration, management practices, tile drainage density, etc.) and which would be the primary entry point for any mitigation measures are not directly considered. Accordingly, the effect of any changes in these features (e.g. management practices) cannot be evaluated (not saying that physically-based models are per se any better with regard to that given the parameter uncertainty). In my opinion, the presented DBM models are much better suited to analyze the effects in changes of the precipitation (as rainfall is the main input) under the condition that these future precipitation conditions are covered in the calibration period.

We agree with the reviewer that DBM models in isolation cannot be used directly to evaluate different mitigation measures, but we believe that the understanding of catchment function, that for DBM models is determined from the data alone, is still helpful in planning appropriate mitigation measures e.g. targeting of fast/surface pathway in the Newby Beck/Eden catchment. Indeed, as experiments on various mitigation measures develop and allow for mapping of the effects of the mitigation measures onto the parameters of the DBM model (e.g. Chappell et al., 2006 and current projects in Lancaster Environment Centre, Dr N Chappell), the potential for using DBM to "help in planning appropriate pollution mitigation measures" will increase. Unlike physically-based models, in which the (generally unknown) structure is fixed beforehand, with parameters optimised to make it fit the data, the dominant modes of catchment function determined from DBM models can be interpreted directly. However, we propose to modify the abstract and text to say "The models led to a better understanding of the dominant transfer modes, which will be helpful in determining phosphorus transfers following changes in precipitation patterns in the future."

Chappell, N. A., Tych, W., Chotai, A., Bidin, K., Sinunc, W., and Chiew, T. H.: BARUMODEL: Combined Data Based Mechanistic models of runoff response in a managed rainforest catchment, Forest Ecol. Manag., 224, 58-80, 2006.

In addition, I would love to see some more analysis of the very nice data they collected. I would assume that the manuscript would greatly benefit if the model results would be discussed together with the data (for example detailed analyses of the hysteresis curves).

We did not include detailed analysis of the high-frequency data as this has already been published by several authors, e.g. Outram et al., 2014, HESS (including hysteresis analysis); Perks et al., 2015, Sci. Tot. Environ; Ockenden et al., 2016, Sci. Tot. Environ.

Specific remarks:
Title: Improvement compared to what, other models? **Yes, this was compared to other models, but in a general sense only, as direct comparison is not possible unless on the same catchment with the same dataset. However, we propose to revise the title to "Prediction of storm transfers and**

annual catchment phosphorus loads with data-based mechanistic models using high-frequency data".

P1, L31-32: See comments above

See response above

P2, L7: The authors correctly point out the importance of the measurement uncertainty. However, in the whole manuscript no information is provided regarding the uncertainty of the rainfall, discharge and phosphorus measurements or how this uncertainty is handled in the modelling approach. Especially the stage-discharge relationship (regarding the discharge measurements of flood events) can be subject to considerable uncertainty which would directly translate into uncertainty of the phosphorus loads. One could argue that the measurement uncertainty is indirectly accounted for by the parameter uncertainty. However, given that the uncertainty bands are hardly detectable in the figures and measurements (without error bars) are not covered by it, it seems that either an important process is not captured by the model or that the measurement uncertainty is underestimated.

The figures currently show only the uncertainty resulting from parameter estimation, and with good model fit, that is usually small. We propose to show 'double banded' plots with one band on the observations to show measurement uncertainty on the discharge and phosphorus load, and one band on the model simulation to show model parametric uncertainty.

P2, L24-32: Here, the authors report the disadvantages and shortcomings of large, overparameterized process-based models (e.g. SWAT). I understand the motivation for that and even to a large degree agree with them. However, the authors should not only pick and describe the most extreme (or worst) process-based models. There are also parsimonious process-based models which can deliver reasonable results describing dynamics of phosphorus on hourly time steps (for example Hahn et al., 2013) or spatial herbicides losses (which have very similar transport pathways) (for example Frey et al., 2011) with few parameters.

We have recognised that there is a wide range of models of differing complexity (p2, l17-24) which are applicable in different circumstances. We wanted to contrast the two ends of the scale, which is why we picked on the SWAT model. However, we propose to include some further examples of more parsimonious models as mentioned above: "Less complex process-based models, with fewer parameters, have also been developed for phosphorus transfer and have been applied with reasonable success to specific catchments, e.g. Hahn et al., 2013; Dupas et al., 2016. Both these studies related to small catchments (< 10 km^2); it was recognised that the models would only be applicable to locations where the assumptions of the model were satisfied, which is consistent with Beven (2000) and 'uniqueness of place'."

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, 4, 203-213, 10.5194/hess-4-203-2000, 2000.

P4, L14: What was the motivation to measure TP and not distinguish between or focus on particulate and/or dissolved phosphorus? Particulate (PP) and dissolved phosphorus (DP) can have different pathways and dynamics. While PP often shows a clockwise hysteresis (P peak before Q peak), DP often shows an anti-clockwise hysteresis (Q peak before P peak) (Dupas et al, 2015). By modelling them separately, it would be probably easier to identify a suitable transfer function and the corresponding pathways.

This is a fair point, and ideally we would have looked at both TP and DP/PP. However, we did not have the data for dissolved phosphorus, as none of the bank-side analysis was done on filtered samples. Both TP and Total Reactive Phosphorus (TRP) were measured by the Demonstration Test Catchments teams, who collected the data. It would be interesting to model TRP too (which could be used as an approximation to dissolved reactive P, but we concentrated on TP in this study as the ultimate goal (for the NUTCAT 2050 project, of which this study was part) was to predict TP loads under climate change.

P5, L17: What is R in the equation, rainfall?

Yes, it is defined on p5, l2.

P6, L6-10: What is the motivation for setting up these short-term models for the Newby Beck catchment when the long-term model have similar performances and structures?

The short-term models were linear, i.e. an even more simple structure and even fewer parameters. The purpose was to show that for short periods, where antecedent flows for events were rather similar, a model with just five parameters could be identified. We are evaluating a methodology in this paper, and successful modelling at different time scales can be used as a validation of the approach. This is particularly the case when validating over extreme events – even given the large uncertainty of discharge observations during the selected period (Storm Desmond).

P6, L15: If I understand it correctly, the output which is used to identify and calibrate the model is also used as an input. I find this contra-intuitive and not really "proper". Why not use a precipitation based antecedent wetness index?

APIs (antecedent precipitation indices) introduce additional parameterisation, often arbitrary, which is exactly what we are largely avoiding by using DBM methodology. Using a simple nonlinear function (with a single and optimised parameter) of recent discharge measurement as catchment wetness surrogate has been tested on catchments of different size and nature, and published numerous times (e.g. Young and Beven, 1994; Chappell et al., 1999; Young, 2003; McIntyre and Marshall, 2010; Beven, 2012). After all, a recent high flow will be highly correlated with high 'overall' catchment wetness, and using the flow at time t-1, as in Eqn. 4, still allows estimation of Re and Q at time t. A simple antecedent precipitation index was actually tried; it worked with reasonable success for Newby Beck but not for the other catchments, and therefore, as a consistent method was sought for all catchments, the API approach was not pursued in this case. We propose to mention this approach in the manuscript.

Young, P. C., and Beven, K. J.: Data-Based Mechanistic Modelling and the Rainfall-Flow Nonlinearity, Environmetrics, 5, 335-363, 1994.

Chappell, N. A., McKenna, P., Bidin, K., Douglas, I., and Walsh, R. P. D.: Parsimonious modelling of water and suspended sediment flux from nested catchments affected by selective tropical forestry, 354, 1831-1846, 10.1098/rstb.1999.0525, 1999.

Young, P.: Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale, Hydrol. Process., 17, 2195-2217, 10.1002/hyp.1328, 2003.

McIntyre, N., and Marshall, M.: Identification of rural land management signals in runoff response, 24, 3521-3534, 10.1002/hyp.7774, 2010.

Beven, K. J.: Rainfall-runoff modelling : the primer, 2nd edition, John Wiley & Sons, Chichester, 2012.

P7, L7-10: Some scatter plots would be very helpful to illustrate the Q-P relationships.

Discharge-TP concentration plots for the three catchments are already given in Supplementary Information Figures S1 – S3. However, we propose to refer to them here as well.

P7, L16: Table number is missing.

Thanks for noticing. It should be Table 1.

P7, L19-20: Because Blackwater has the lowest specific discharge. It would be good to discuss and explain the differences in the specific discharges and P concentrations among the catchments.

Agreed. We propose to add a column in Table 1 giving the rainfall-runoff ratio for that year, and to change text to say "The lowest mean annual TP concentrations were observed in the Newby Beck catchment, but combined with the highest runoff this resulted in a high total annual TP load. Conversely, although mean annual TP concentration in the Blackwater was also higher than in Newby Beck, when combined with the lowest runoff, this resulted in the lowest total annual TP load. The rainfall-runoff ratio for Newby Beck (0.65) was much higher than for the Blackwater (0.30) or the Wylye (0.32), indicating a larger capacity for storage in the latter two catchments.

Despite similarity in the rainfall-runoff ratio, total runoff in the Wylye was higher than the Blackwater because of the higher total rainfall.”

Differences in the P concentrations are already explained in the paragraph p7, l3-15.

P7, L28-30: So were model results actually used to fill data gaps for the longterm model? If yes, this should be clearly stated and discussed accordingly.

The linear model would only have been used to fill data gaps in the short-term data series, if a complete series was required to estimate, for example, TP load over the calibration period, based on observations. This was not actually used in this study. However, the DBM transfer function models can be used in model-based interpolation of the output, given the input signals, just as they can be used in forecasting (e.g. Smith et al, 2014).

Smith, P. J., Panziera, L., and Beven, K. J.: Forecasting flash floods using data-based mechanistic models and NORA radar rainfall forecasts, 59, 1403-1417, 10.1080/02626667.2013.842647, 2014.

P7, L29-30: How can model results help in identifying problems in the extrapolation of the stage-discharge relationships, when the whole model itself is based and calibrated with data of exactly these stage-discharge relationships? In my opinion the model is only reliable for the conditions covered during the calibration period. If more extreme events would be included in the calibration period, the model and the parameters would very likely be different.

We did not mean to imply that the model could identify problems with stage-discharge relationships, but rather to suggest that it could be useful when there are known problems with the relationship. However, we agree with the comments about model validity outside of the calibration conditions. The calibration (in the sense of stage to discharge) is the weakest point of all the catchment models relying on stage measurements, particularly for extreme events.

P8, L4: Should be “Table 2”

This will be changed.

P8, L4-18: I find the discussion and evaluation of storm Desmond a bit constructed and unnecessary. You don’t need a DBM model to realize that discharge and P load was underestimated when there are reports of out-of-bank discharge bypassing the gauging station. The model also doesn’t help in the quantification of the missed P and Q. As mentioned before, the model was trained under different conditions and is therefore in my opinion not really valid for very extreme cases not being part of the calibration period (again not saying that physically based models are any better).

We agree that the results for Storm D are tentative, they are shown here to demonstrate the effectiveness of DBM over a particularly challenging period in data.

Table 2: According to the time constants and order of the Q- and TP models, there are two pathways contributing to the discharge generation with only the fast pathway contributing to the TP generation. If I understand the concept of the TC correctly, TP reacts before the discharge rises. Is this in agreement with the measured data?

Shorter time constant in the case of impulse shaped input means that the response grows faster and decays faster, not that it reacts quicker (that would be the time delay, which in this case is the same).

Table 3: What is the meaning of the term “using Qsim”. If model outputs instead of actual measurements were used, this should be clearly stated, justified and discussed (for example why is the performance worse for “using Qsim” that “using Qobs”?) In relation to that, how was TPLoad calculated? Did the authors used the modeled Q to calculate TPLoad or did they use the measured Q? Again, if modeled Q was used, this should be stated, justified and the consequences discussed.

Thanks for pointing this out – we have not expressed it clearly. The effective rainfall is calculated according to Eqn. 4, using the observed discharge, Qobs, as a proxy for the storage state of the catchment. Model parameters for the linear model (effective rainfall-runoff) are estimated from this. This results in Rt2 using Qobs. However, for a true simulation, Qsim is calculated only from

the rainfall and the model parameters, giving Rt2 using Qsim. The effective rainfall – TPload model is a two-stage model; it is assumed that the discharge is unknown, so that the effective rainfall must be calculated one step at a time, as Qsim is generated with the previously identified parameters of the rainfall-discharge model. Hence Rt2 using Qobs is a one-step ahead prediction, whereas Rt2 using Qsim is a true simulation, only using the rainfall input.

TPload was calculated according to Eqn. 1, using the observed discharge and the observed concentration.

Table 3 cont'd: For the Newby and Wylye TPLoad models effective rainfall was used as input, while regular rainfall was used for the discharge model. What is the meaning of that? Does it mean that for TP dynamics antecedent conditions are important, while they are not important for the discharge dynamics? Again, I would advise to discuss these findings as well as the different time constants and their percentages together with the actual measured data.

Thanks, we realise that our explanation is unclear. All models, apart from the Blackwater rainfall-TP model, are linear models using effective rainfall as input. The effective rainfall is calculated using a non-linear function, according to Eqn. 4. The antecedent conditions are important in both discharge and TP dynamics. The reason the effective rainfall was not used in the Blackwater TPload model is because the simulated discharge, Qsim, is a poor fit (Rt2 using Qsim = 0.37, which is worse than for a rainfall-runoff model with linear rainfall input). We propose to change Table 3 to make it clear that effective rainfall was used in all cases except the Blackwater TPload model

P9, L26: What does "effective rainfall (from the runoff model)" mean?

This means effective rainfall calculated one step at a time using Qsim.

P11, L7: Same point again. What does "effective rainfall simulated by the rainfall-runoff model mean?

as above

P12, L1: It's nice to have models with a low parameter uncertainty. However, when the uncertainty bands do not encompass the measurements, it's not really better situation than having a large parameter uncertainty. The model is either missing an important process or measurement uncertainty is not accounted for. A third reason could be a too narrow parameter sampling space in the MC method.

Parameter sampling used in the MC runs is from a multivariable Gaussian distribution using the estimated parameter values as means and their estimated covariance matrix as covariance. The model fits the data well, so the covariance matrix is small (in L2 sense), and the uncertainty of the model is limited to its parametric uncertainty. What is not accounted for here is the uncertainty of the measurements – see response above. We propose to show figures with double bands – a band on the observations, indicating the measurement uncertainty, and a band on the model simulation, indicating the parameter uncertainty. This will bring the additional value of visual partitioning the uncertainty of model predictions. Thanks for pointing this apparent issue out.

P12, L28-33: Although, the authors openly discuss the limitations of their modelling approach, there is one point I miss. They argue that understanding the rainfall-Q/TPLoad relationship through DBM models can help to identify dominant modes of the catchment and can therefore be used to target management interventions. I would argue that this is only possible if the identified dominant modes or pathways can be related to specific areas in the catchment. In my opinion it is not enough to know that 70% of the TPLoad was activated via a fast pathway. It is necessary to know which areas in the catchments are connected to the stream via this pathway, how these areas are managed and what their soil P status is. To actually plan and implement intervention strategy, you need to know where (on which fields) and how to intervene. The "how" is strongly dependent on the "where". If you identified some fields with subsurface tile drainage as the contributing areas you would need a different intervention strategy as for example on a field with a tendency for surface runoff due to soil compaction. Knowing the temporal dynamics is not good enough, you would also need information about the spatial patterns.

This is true of any model where the observations are from the catchment outlet. It is not possible with any certainty going beyond assertion to apportion the contribution of specific areas without observations characterising these specific areas. We accept that DBM does not provide information about the spatial patterns, but we did not claim that it could be used to "target" management interventions (this makes it sound location specific), merely to be useful in "planning" interventions. We propose to tone down the text in this respect, modifying the abstract to say "The models led to a better understanding of the dominant transfer modes, which will be helpful in determining phosphorus transfers following changes in precipitation patterns in the future."

Authorship:

I thought long about including this very last comment in the review. However, given the many discussions I had with colleagues about this very issue in the past, I feel somewhat obliged to mention that I find the number of authors contributing to this manuscript too excessive, given the nature of the article (a regular modelling study). I am very much in favor in acknowledging significant contributions (for example with respect to data gathering) with a co-authorship, however this seems not to be the case here. The authors state themselves that while two persons were responsible for the modelling, three persons did project management and the remaining fourteen (!) basically discussed the results and did some editing. I certainly don't want to offend any of the authors and obviously have no insights in the preparation process of the manuscript. Nonetheless, I would encourage each co-author to reflect if in their opinion they really contributed significantly to this manuscript.

This modelling study contributed to a large consortium project (NUTCAT 2050), the ultimate aim of which was to make predictions of phosphorus transfer into the future. As part of that project, this modelling was developed and discussed with the project team, alongside other modelling approaches. All members of the NUTCAT 2050 team have been involved in the evolvement of this modelling study and have contributed to the manuscript. Other co-authors were involved with the Demonstration Test Catchment (DTC) Project, which collected the data. To clarify, we propose to add to the author contributions "MCO, KJB, ALC, RE, PDF, KJF, KMH, MJH, RK, CJAM, MLV, CW, PJW, JGZ and PMH contributed to NUTCAT 2050; ALC, KMH, SB, RJC, JEF and PMH are part of the DTC project."

References

Dupas, R., Gascuel-Odoux, C., Gilliet, N., Grimaldi, C., Gruau, G., 2015. Distinct export dynamics for dissolved and particulate phosphorus reveal independent transport mechanisms in an arable headwater catchment. Hydrol. Process. 29 (14), 3162e3178. http://dx.doi.org/10.1002/hyp.10432.

Frey,M.P.; Stamm,C.; Schneider,M.K.; Reichert,P. (2011) Using discharge data to reduce structural deficits in a hydrological model with a Bayesian inference approach and the implications for the prediction of critical source areas, Water Resources Research, 47(12), W12529 (18 pp).

Hahn,C.; Prasuhn,V.; Stamm,C.; Lazzarotto,P.; Evangelou,M.W.H.; Schulin,R. (2013) Prediction of dissolved reactive phosphorus losses from small agricultural catchments: Calibration and validation of a parsimonious model, Hydrology and Earth System Sciences, 17(10), 3679-3693.