Authors' response to Reviewer 4, Anonymous
For clarity, we have included the reviewer's comments in black; our response is in blue

General comments
The paper is well written and shows an interesting analysis of an extraordinary dataset with sophisticated modeling tools.
Thank you

Besides these merits, the manuscript would clearly benefit from a new title, a somewhat more balanced judgement on data-based models, a clearer model introduction and a better separation from Ockenden et al. (in press).
See later comment on Ockenden et al. (now published)

The title promises improving "predictions of storm transfers and annual loads in surface waters", yet the outcome was rather "a better understanding of the dominant nutrient transfer modes, which will, in-turn, help in planning appropriate pollution mitigation measures" (last sentence of abstract).
We accept the comments on the comparison with other models and we propose to revise the title to "Prediction of storm transfers and annual loads with data-based mechanistic models using high-frequency data".

I agree more, but not completely, with the abstract: the manuscript is about transfer modes derived from observed data and not better predictions (the predictive power of the models was rather limited). Loads did not get too much emphasis either. Therefore, I suggest to change the title to match the main findings.
We accept the comments on the limitations of the model, but we maintain that the models have very good predictive power where there have been no fundamental changes to the catchment, and under similar input conditions. We propose to change the title as above. We propose to tone down 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."

According to the authors, a major advantage of the data-based approach is that rather complex models can be used to predict Q and TP without the need to know anything about the major processes – knowledge will be extracted from the data.
The advantage of the data-based approach is that rather simple models can be used. The advantages and limitations of the DBM modelling approach are listed in Table 4.

Indeed, complexity of their model rivals that of certain conceptual models. However, this application wasn't a clear success. The intro and the conclusions are very optimistic about databased models. However, such DBMs have the same problems as conceptual models: their elements are so abstract that they can't be linked to anything observable. This semi-black-box behaviour is nicely demonstrated in the manuscript: a strong 'slow' component is present in TP in certain catchments, various possible reasons are listed, but there is no way to know which applies in the specific case (e.g. SRP in baseflow may indeed come from WWTPs or activated deposits, etc, but which case does apply here? [We are not informed whether there are WWTPs in the catchments.]). So while process-based models are typically overparameterised and laden with uncertainty, their less abstract formulation leaves open at least theoretically to gather additional observations to prove or falsify hypotheses.
We accept that there are limitations to what you can interpret from DBM models – such as the spatial differences within a catchment. But we maintain that these issues (such as the WWTP

sources mentioned above) are exactly the same as with process-models, except that for those models you have to hypothesise first in order to decide what processes to include in the model. There is also a danger in process-based models that one could draw conclusions about the components of the system dynamics not present in the data (and so not indentifiable), thus using model artefacts (or noise patterns within uncertainty) to draw such conclusions. This assists in getting the right answers for the right reasons (Kirchner, 2006).

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, 42, W03S04, doi:10.1029/2005wr004362, 2006.

Overall, I think that it should be mentioned that DBMs pay with an extreme data demand for not asking a priori knowledge on the system. Given that unresolvable issues appeared with such an extremely good data coverage, it is somewhat dissonant to recommend DBMs for catchment management when (i) model constituents can't be linked to anything else than the input, and (ii) there are practically no other catchments in the World with a comparable data coverage.

The high data demand of DBM models is noted in Table 4 and the section on advantages/limitations of the modelling method. The DBM transfer function (TF) models link the observations of inputs and outputs in a causal structure where the model is a system of differential equations (normally, but not necessarily linear) picking up the dominant dynamic modes present in the data. This is not a black box model, as the model structure is selected in the modelling process as a physically feasible one (hence the Mechanistic in the name) and identified from the data, which eliminates the modes of behaviour that are not present in the data, and thus not identifiable (instead contributing to the uncertainty). The basic component of TF models – the well known ADZ model for flow-flow models but also for rainfall-runoff models (Young, Beven, others) is based on mass balance between the input and output signals (flow, rainfall), the structure of such mass transfer blocks connection (parallel, serial) is also based on physical considerations, just based on other equations/paradigms compared to physics/process based models.

Just because data coverage in many catchments is poor does not mean it should always stay that way. Technology and monitoring methods are improving all the time so that high-frequency data are now more readily available, e.g. Jordan et al., 2005, 2007; Outram et al., 2014; Skeffington et al., 2015. We should embrace efforts to improve data coverage and ways to use it wisely.

Jordan, P., Arnscheidt, J., McGrogan, H., and McCormick, S.: High-resolution phosphorus transfers at the catchment scale: the hidden importance of non-storm transfers, Hydrol. Earth Syst. Sci., 9, 685-691, 2005.

Jordan, P., Arnscheidt, A., McGrogan, H., and McCormick, S.: Characterising phosphorus transfers in rural catchments using a continuous bank-side analyser, Hydrol. Earth Syst. Sci., 11, 372-381, 2007.

Outram, F. N., Lloyd, C. E. M., Jonczyk, J., Benskin, C. M. H., Grant, F., Perks, M. T., Deasy, C., Burke, S. P., Collins, A. L., Freer, J., Haygarth, P. M., Hiscock, K. M., Johnes, P. J., and Lovett, A. L.: High-frequency monitoring of nitrogen and phosphorus response in three rural catchments to the end of the 2011-2012 drought in England, Hydrol. Earth Syst. Sci., 18, 3429-3448, 10.5194/hess-18-3429-2014, 2014.

Skeffington, R. A., Halliday, S. J., Wade, A. J., Bowes, M. J., and Loewenthal, M.: Using high-frequency water quality data to assess sampling strategies for the EU Water Framework Directive, 19, 2491-2504, 10.5194/hess-19-2491-2015, 2015.

Another issue is related to the model presentation and the relation to the paper by Ockenden et al. (in press). The Methods section provides a very brief overview of the models referring to the companion paper for details and for calibration, so the models were first published and calibrated therein (this can't be verified because the article isn't accessible at the moment, it's still in press). Please highlight the novel parts of this manuscript compared to Ockenden et al. (in press).

This paper is a companion paper to Ockenden et al. (Nature Comms, 2017), now published. That paper uses a DBM model as part of a multi-model study to predict phosphorus transfers in the future, i.e. an application of the model developed in this paper. This paper provides full details of calibration and validation and, particularly, interpretation of results, which is not included in Ockenden et al., 2017. However, we propose to expand the methods section here to include full details of the models.

If the main novelty is the dynamics of TP load, the analysis of the results could be somewhat extended at the cost of details on the fast/slow components of TP (which are presented to the last detail). For example, it would be useful to elaborate more on the TPflux vs Q relationship. Yes, Q was already used to calculate TPflux, but the final correlation is actually determined by the relative variance of Q and TPconc. This would highlight how much delay and nonlinearity (both causing hysteresis) is present and therefore how much we gain by having a nonlinear autoregressive model.
There two types of lags: pure time delay (the 'Poohstick time' for the flow-flow models) and dynamic lag resulting from mass transfer dynamics, both contributing to the observed hysteresis loop but resulting from different phenomena, not easily distinguishable from the shape of hysteresis, but perfectly identifiable and quantifiable in the DBM/TF approach. Relative variance? That would indicate a static relationship between Q and TPconc. What we established from the data is that this relationship is dynamic.

Independent of this, deriving the applied models from the general 2nd order continuous transfer function (TF) model seems to be an unnecessarily complicated choice for several reasons:

• Only those models were accepted, which could be converted to the parallel linear storage format **(or serial, or indeed first order depending on the structure identification results).**

This is very welcome, because such models are Markovian, i.e. the system's current internal state (or the last state in discrete formulation) and the current inputs completely determine the system's response. This assumption is typically made in most environmental and hydrological models. In contrast, a 2nd order TF model can be non-Markovian too (=a long system history is required to understand the current response, actual state is not a complete descriptor), which would be very hard to justify
Any order TF model has a range of equivalent state space Markov type models, which is clearer in their discrete form, but the discrete and continuous forms of TF models are equivalent. We are not sure what the Reviewer is referring to here.

(which physical/biological/chemical process would lead to such system? – **any uncertain mixing or transport process).**

• Continuity dominates in the model description, while eq. 4 and the aggregation to 30 minutes make it obvious that inputs were treated discretely
Continuous time models are estimated from sampled data – there is no contradiction here.
Then why bother with the more complex continuous models?
Continuous time models are more numerically robust and have a direct interpretation as systems of differential equations (Young, 2011)
Young, P. C.: Recursive Estimation and Time-Series Analysis: An Introduction for the student and practitioner, Second ed., Springer, New York, 504 pp., 2011.

These kinds of models are not used very frequently in hydrology/water quality modeling (as opposed to signal processing)

They have not been used widely until recently because of the lack of effective model identification methods such as those in the CAPTAIN Toolbox – this is a part of the novelty of this work.

Potential readers may easier understand if 2 parallel linear storages
or ARX models were mentioned as alternative formulations for the same model.
This is the case for discrete time models, which have no direct differential equation interpretation and are less robust with respect to stiff systems with dynamic modes of very different dynamics (Young, 2011).
Young, P. C.: Recursive Estimation and Time-Series Analysis: An Introduction for the student and practitioner, Second ed., Springer, New York, 504 pp., 2011.

• If the parallel storage formulation is so important to learn about slow and fast
components, why are parameters shown in the general 2nd order form in Table S5?
Table S5 shows the general polynomial form for information (and in case anyone wants to simulate using this model), because they are estimated in this form, so their parameter uncertainties are obtained in the general polynomial form too. Factorisation of the rational polynomial TF into parallel, serial etc. components is the next step in the DBM process. The decomposed form, with time constants for slow and fast components are given in Table 3.

Smaller issues
Page 2 Line 25: USLE is more semi-empirical than process-based.
Accepted. This will be changed

Page 4 Equation 1: Use consistent units. If Q had [m3/h], and TPconc [kg/m3], TPload
would readily be in [kg/h] without a conversion constant.
This is true, but would then result in working with values that are either very large (in the case of Q) or very small (in the case of TPconc). We felt that it was better to stick to the units in which the variables were measured.

Page 5: Were the tau (delta) delay constants calibrated or fixed?
Estimated from data using information criteria, just as for the model structures.

Page 6 L 15: If Re was necessary because the internal state of the catchment affected
runoff and TP transport, would it make sense to use Re to the model Q as well?
Re was used in all models except the Blackwater rainfall-TPload model. However, we accept that this does not come across clearly in the text. Q is modelled as a linear transfer function with Re as input, where the non-linear relationship between R and Re is estimated at the same time as the TF parameter estimation. TPload is also modelled as a linear transfer function with Re as input, except that Re (and Q) in this case are first simulated using the parameters previously estimated for the R-Q model.

Page 7 L 22-25: Of course, most pollutants do not follow Q, because they have either limited or temporarily activated sources or they partition between water and sediment. According to your argument on celerity, even non-partitioning conservative pollutants would theoretically show a hysteresis.
A truly 'non-partitioning conservative pollutant' would act as a conservative tracer, moving exactly with the water particles (which forms the basis for dilution gauging using conservative tracers). In this case there would be no hysteresis on the phase-plot between Q and that hypothetical tracer. This does not contradict our comment on the hydrograph representing the integrated effects of celerities.

Page 9 L 3-4: Converting these constants to half-lives would make them easier to judge. It is somewhat difficult to grasp decay to 1/exp(1).

The half-life interpretation is only a good illustration for pure recession curves (with the ln(2) proportion between T_C and $T_{1/2}$), it is less obvious for differential equations with changing complicated inputs. This is the standard definition of a time constant in a first order linear time-invariant dynamic process e.g. $A(t) = A_0 \exp(-t/T_c)$ where T_c is the time constant, commonly used in hydrological literature (see Nash cascade definition, Shaw et al, 1994 etc).

Shaw, E. M.: Hydrology in Practice (3rd edition), Chapman and Hall, London and New York, 1994.

Page 10 L 18-21: If we don't know the mechanisms responsible for the slow pathway, what kind of measures could be taken?

These models do not provide recipes or final answers, but objectively point out specific parts of the system dynamics. Having identified that a slow pathway is so important, measures which prevent pollutants getting to the slow pathway in the first place, such as reductions at source, will be helpful. This may require further specific measurements, such as testing P in soils or identifying septic tanks in the catchment. The difference between DBM and process based models is that this interpretation in DBM models is made a posteriori, after the data assimilation and is based on objectively identified quantitative features of the process, with process based models the interpretation is done a priori, with all the caveats related to such a sequence.

Page 12 L 24-28: It's true that process-based models make some assumptions that do not always hold, but here it was demonstrated that neither the DBM can always be validated. Time-variable parameters are a useful concept, but seldom implemented. As fluxes of TP are modelled, Fig S1 should rather show TP fluxes against Q. This could reduce clutter and illustrate how a naive linear model would work. Considering my comment above, it would be useful to move a modified version of this figure into the main text.

The plots of Q against concentration are shown to illustrate the hysteresis loops and to show the background concentration at varying baseflows. However, we propose to add plots of Q against TPload to the supplementary information but do not feel that this would add to the main text. A linear dynamic model does not have to be naïve if it is identified from the data using a rigorous procedure. It indicates that this is the maximum model approximating the data well, that can be estimated based on the present data set.

A major model-based finding of this study is the demonstration of are the importances of fast and slow pathways of Q and TP in the catchments. This could, at least partially, be derived directly from the data! High baseflow indices and slower recession indicate important slow pathways of Q, high baseline concentration indicates the same for TP.

There are many, often very arbitrary methods of base flow estimation and there is much discussion in the literature as to which one is better, such heuristic dominance 'approximation' has neither rigorous quantification nor uncertainty estimation elements, unlike the DBM modelling procedures.

As the applied model doesn't have any mechanistic explanatory power (e.g. identification of reasons for these), how could management benefit from modeling? Please comment on this briefly.

See response as for Page 10 L 18-21 query.

On load figures, load is [kg], but per which time unit?

All units are hours. Load figures will be changed to kg h^{-1}.

What are the time units in e.g. Table 2? If the continuous version was used, time has

to have a unit. If the discrete version was used, the applied timestep has to be written.
Time units are hours. This will be added to Table 2 (and Table 3).

What are the other units in Table 2?
Σobs and Σmodel are totals for the period of Storm Desmond. These are in mm for runoff and kg for TPload. These will be added to Table 2.