Thank you for your feedback and comments to our manuscript.

The conclusions of the paper (section 5) unfortunately do not go much beyond what was already known prior to model application (the same is true for section 4.2 on model limitations)
We only partially agree with this critique. We do think that this study resulted in interesting findings that give new insight and have not been fully outlined before. On the other hand, we agree that sections like the Abstract or the Conclusions convey this information only to a limited degree because they contain (in their current form) too many generalities for which the assessment above holds true. Below we list explicitly the main points we consider the relevant findings. Upon revision of the manuscript, we will point out these issues more clearly.

> 1) The three approaches represent (partially) different influencing factors: soil type, topography and connectivity. It is known from the literature that each of these three factors may play a crucial role for diffuse P losses to surface waters. However, comparisons of approaches based on them in different ways are not wide-spread. Accordingly, the outcome of the comparison is not evident before the analysis.
> 2) The comparison of the RRP and the SCIMAP model shows how the risk predicted by SCIMAP may vary as a function of event size (change in time) and of soil type (spatial heterogeneity) (see Fig. 7). Such analyses have been lacking so far. SCIMAP identifies similar critical areas for DRP without using time series data. Hence, this information can be used to extend SCIMAP as a valuable screening tool.
> 3) While Lane et al. 2009 also evaluated the Network Index using a dynamic model they suggest that it is necessary to further investigate the index's potential with regards to the duration of integration (monthly, yearly, decadally). Our finding that the stepwise linear relationship with zero risk up to the 5 % NI quantile and a maximum risk level with no further change at the 95 % Network Index (NI) quantile was appropriate for storm events, is in contrast to previous expectations (Lane et al., 2006; Lane et al. 2009): To delineate the connection risk from the network index we used the approach suggested by Reaney et al. 2011. A connection of 0 was assigned to places with NI values below their 5 % quantile and a connection of 1 to places with NI values above the 95 % quantile. Between the 5 and 95% quantiles a linear relationship between NI and the connection risk was assumed. Our study showed that this approach produces model results similar to results from the RRP model for storm events. It has previously been assumed that SCIMAP predicts the average risk over years rather than over an individual storm event. For average risk predictions we suggest a modification of the relationship between NI and connection risk.
> 4) Furthermore, we point out that SCIMAP needs some rescaling if it is used for comparative purposes across catchments.

The study is further limited methodologically by using the RRP model as a benchmark to compare the other models against. The authors justify this based on their calibration/validated of the model as reported in Hahn et al., 2013). However, the calibration timeseries was extremely short (7-17 July 2000, P14504, L5) and I doubt that all important modes of behaviour of the system are reflected in the calibration information and hence the model parameters. The validation periods used in the present paper are equally short (Mar-Nov 1999 and May-Aug 2010 for the 2 catchments, respectively). So RRP is, in my opinion, not a robust benchmark.
The Reviewer raises a very important issue: what is a meaningful benchmark for method comparison? Obviously, direct, spatially distributed flux measurements of DRP (reaching the stream!) would be optimal. However, there is no such data that can be used as a benchmark. This is a fundamental limitation of most studies on distributed hydrological model and we cannot solve this problem here. As a substitute, we argue that the predictions from the RRP model are a useful benchmark for our purpose because i) the RRP model is the most complete of the three approaches accounting for most of the factors included in the others, and ii) because the RRP model was

validated using not only data from the catchment outlet but also spatial data (see p. 14499, L. 56 -7). Furthermore, we consider that the calibration procedure performs better than the reviewer suggests:

Short calibration time series:

- The observation is correct that the calibration period was short. However, the crucial question is whether the calibration period covered a sufficient range of the relevant system states during this period. As mentioned earlier (Lazzarotto et al., 2006) and restated in our recent publication on the RRP model (Hahn et al., 2013) this calibration period covered a wide range of streamflow conditions. This aspect is more important than the length of the calibration period (Gupta and Sorooshian, 1985; Lazzarotto et al., 2006; Yapo et al., 1996). We will briefly mention that in the manuscript.

Validation time series:

- The main DRP losses occur during the growing season. Accordingly, the model includes processes relevant during this period but neglects for example snow cover and snow melt. Thus the model should only be used for periods between March and November. The first validation period covers this time span, while the validation period in 2010 is indeed relatively short. Still it is about eight times longer than the calibration period and covers a wide range of soil moisture and hydrological conditions (see (Hahn et al., 2013)).

Based on these arguments we consider that the calibrated RRP is a reasonable benchmark. However, we should make it more explicit in the discussion that the benchmark is not real data but the model output. This needs to be considered when drawing conclusions from the results. We will check the wording in the manuscript to avoid making too bold statements in this regard.

The authors state regarding the validation of the DoRP model that "no reliable statement for the Stägbach catchment is possible due to the limited number of observations" (P14505, L17-18), and I believe the same is true for the other catchment and for RRP.
We cannot follow the argument why the limitation that we state for the Stägbach is generalized by the reviewer to the catchment and RRP in general:

- This statement refers to the fact, that only three runoff events were used to assess the performance of DoRP in the Stägbach catchment. Thus, it is not clear whether we see a trend or whether the one data point with high discharge in Fig. 3 is an outlier.
- For the Lippenrütibach catchment more information was available
- The RRP model can simulate the whole time series, not only the discharge for certain events. Thus, enough data points for validation are available.

In addition, I had the following comments:
P14498, L28-29: Is this not pre-empting the results? What about DoRP?
No, here we just wanted to point out that: DoRP itself only comprises the hydrological part. RRP and SCIMAP on the other hand already include a hydrological and a source (or phosphorus) part. We'll modify the manuscript to make this clear. In addition, the three approaches account for different influencing factors (see above) and it is not clear from the beginning what the outcome of this comparison will be.

P14500, L7: Please explain uniform MC method.
P14500, L7: "The model was simultaneously calibrated (uniform Mote Carlo method) on discharge data from four catchments draining into Lake Sempach."

The four catchments varied in their soil composition and hydrological response. The model parameters were determined by repeated random sampling from a uniform prior distribution within the range of each parameter. The performance of each parameter combination was assessed by comparing simulated discharge with measured discharge in the four catchments.

P14500, L9: "Using the modified Nash-Sutcliffe criterion NSC as defined by Lazzarotto et al. (2006) and a NSC threshold of 0,6 724 parameter sets out of 5 million were judged behavioural and used for model application"

P14500, L9-12: Is this not the classic GLUE method? What is the justification of the choice of performance measure (NSC) and behavioural threshold (0.6), particularly in relation to more sophisticated methods such as formal Bayesian methods and extended GLUE (e.g. Romanowicz & Beven, 2003; Rankinen et al., 2006; Winsemius et al., 2009; Krueger et al., 2012)? This is not discussed in in the original modelling study (Hahn et al, 2013)
This argument is not clear to us: Based on the references given, several quite different issues could be raised. Accordingly, our response addresses several issues:

i) Is it a classical GLUE method? The answer is a partial YES. As in GLUE, we define (in a subjective manner) a threshold for behavioral results. However, we avoid interpreting the relative frequency of behavioral parameter sets causing for example fast flow in a probabilistic sense (see (Hahn et al., 2013)).
ii) Justification for NSC: This critique is a bit surprising given the fact that the reviewer refers for example to (Rankinen et al., 2006). These authors actually rely on NSC for their distinction between behavioral and non-behavioral parameter sets (complemented partially by soft data, see below). Accordingly,
iii) Implicitly, the question by the reviewer suggests that the classical GLUE method is not appropriate for the task described in this manuscript and refers to formal Bayesian methods and to extensions of GLUE. Because, the reviewer only refers to articles on extensions of GLUE we briefly discuss issues emerging from the cited articles that might be relevant in our context. One aspect that is raised by these articles is the inclusion of soft data into the evaluation of parameter sets (Rankinen et al., 2006; Winsemius et al., 2009). In our case, one could have thought about integrating "soft data" like the observation of surface runoff at different locations into the evaluation. However, formulating a well-founded quantitative expression is not straight forward. It seems at least questionable whether a formal inclusion of these data had improved our analysis as compared to our approach to consider these observations for the discussion.

In summary, we consider the approach appropriate to the objective of this paper. We do not see which of the findings is expected to change significantly if one had chosen a different method.

P14506, L23: Homogeneous rather that heterogeneous?
Yes.

P14507, L12-14: I wonder whether the lambda/NI comparison can be given more prominence, perhaps as a new focus of the paper?

Figure SI-1 shows the differences between the topographic index and the network index in our study catchments. Regions with a higher topographic index than network index may not be connected to a stream. The hydrological risk delineated with the RRP model might therefore be overestimated in those areas. The comparison of lambda and NI in our catchments however reveals only minor differences. We therefore think it is more interesting to study the relationship between the RRP risk to generate fast flow and the NI, as we have done in the following sections: page: 14507 27-28 + 14508 (till line 17).

P14507, L21; P14515, L16: Re tile drains, how significant are they in the 2 catchments? If important then topography might not be a good predictor of runoff generation. The same would apply for NI in terms of pollution risk.

The drained area in our catchments amount to approximately 10 % of the agricultural area in the Lippenrütibach catchment and to around 15 % in the Stägbach catchment. For the Cantone Lucerne the drained area is around 11 % of the agricultural area (Unpublished data).

Tile drains draw down the water table, leading to drier soils around the drains than expected based on topography. That means the surface runoff pathway becomes disconnected, but runoff still reaches the stream via the drainage system. In SCIMAP tile drains could be represented directly. Thus, SCIMAP can account for modified soil moisture regimes. The RRP model cannot directly account for tile drains. However, fast flow as defined in the RRP model comprises not only surface runoff, but also sub-surface flow, preferential flow and tile drain flow.

In catchments with a moderate amount of drained area, like our study catchments, topography still provides a good basis for the estimation of runoff generation. Areas where runoff - including tile drain flow - is generated are probably similar, because tile drains were usually installed in very wet places. Thus, runoff is still generated.

Field visits and measurements in the Stägbach catchment showed, that areas predicted to be wet were indeed very wet and surface runoff from some of those areas was registered (even though in one place a tile drain was not very far away). Thus, despite the drainage systems, topography still provides important information about the generation of runoff. The upslope surface area and the slope still determine flow direction and the potential wetness of an area in our study catchments.

P14509, L14-17: Here I'm missing a formal spatial comparison, e.g. via Cohen's kappa.

We based our comparison on Fig. 6 and Fig. 7, which enable a direct comparison.
However, the comment is justified and we now rescaled and grouped the data sets in order to calculate the weighted kappa (J. Cohen, 1968). To rescale the model results we divided the results by the respective maximum value. We then grouped the data as follows:

0 - 0.2 → low risk
0.2 - 0.5 → medium risk
0.5 - 0.8 → high risk
0.8 - 1 → very high risk

The results of the kappa calculation support our statements made on page 14509, L14-17 as well as our findings in section 3.2.2 (The confusion matrices and kappa values are listed at the end of this document):

- The SCIMAP risk predictions are in better agreement with RRP model predictions for a high runoff event (kappa Stäg: 0.54, kappa Lip: 0.68) than for the average DRP load during the simulation period (kapp Stäg: 0.26, kappa Lip: 0.3).
- For average DRP load predictions with the RRP model, kappa is higher if RRP results were compared to the global locational risk (kappa Stäg: 0.29, kappa Lip: 0.45) instead of the original SCIMAP locational risk (kappa Stäg: 0.26, kappa Lip: 0.30).

We will include the results in the manuscript.

P14509, L24-25: I do not think this is a problem since these are the risky times, no?
It is indeed not a problem, but interesting to point out, because so far it was assumed that SCIMAP represents average values

P14510: It is not very clear what was done here – please try and revise.
We revised this section as follows:

The original SCIMAP model prescribes a static linear relationship between NI and the connection risk pcx from 0 at the 5% NI quantile to 1 at the 95% quantile. This relationship is assumed to be time invariant, and to assume that in all catchments the least connected 5% of the catchment never connects while the most connected 5% always connects. This approach has three main limitations. First the comparison with the RRP model shown above suggest that the relationship between NI and connection risk is not time invariant but that SCIMAP predictions mainly reflect larger events in our study areas.

Second, the assumption that 5 % of the catchment is always connected and 5 % is never connected makes the method insensitive to these parts of the catchment. Assuming that areas with very low NI values never connect is reasonable for single runoff events and likely for most monitoring periods. Assuming that areas with the highest 5% of NI values always connect is appropriate for large events, but not necessarily for aggregated risks over a period of time or for small events (Fig. 5b). This can be seen in Fig. 7a and c, which show a considerable scatter for any SCIMAP locational risk. The scatter is reduced when this assumption is relaxed and the connection risk is assumed to scale linearly with NI up to its maximum value (Fig. 8). In our catchment this is because there remain significant differences in connectivity even within the most connected 5% of the landscape. Areas close to the catchment outlet are characterized by very high λ and NI values and, according to the RRP model contribute runoff frequently, even during very small events. However, areas further upstream where the λ and NI values were lower but still within the top 5% contributed runoff less frequently. Extending the linear NI/risk scaling up to the maximum NI enabled differentiation between high NI values and thus between these areas. A third limitation of the original SCIMAP approach is that by normalizing the generation risk and NI values between zero and one it can predict risk only in relative terms within individual catchments. To enable a comparison between catchments, we normalized the generation risk (source factor) and delivery risk (transport factor) by a consistent upper limit for all the catchments. The source factor was divided by the maximum value of all catchments of interest, instead of the maximum for each catchment. Adjusting the transport factor is less straightforward. The highest NI value (NI_{max}) of the two catchments studied here was 20 and the lowest (NI_{min}) was 4.7.The 5 % quantile of all NI values was 6 and based on our RRP model predictions the runoff risk of cells with NI values lower than 6 can be neglected. Thus, we set the transport factor to 0 at NI ≤ 5 % quantile and to 1 at NI ≥ NI_{max} and to vary linearly with NI between these limits. The locational risk calculated with these globally scaled source and transport factors ranged between 0 and 0.4. Using the RRP results as a reference, the global locational risk was in better agreement with the average DRP loads over the whole monitoring period (Fig. 8) than the original locational risk (Fig. 7). Since the catchments had similar soil P status, this improvement can be attributed to the modified relationship between NI and delivery risk.

Tab1/Fig2a: How was the spatial information aggregated over the event timesteps?
For every timestep there were 724 model realizations. For each pixel we counted how many model realizations resulted in fast flow generation. If more than 80 % of the model realizations predicted fast flow generation from that pixel, this pixel was assigned to the very high risk class.

Fig4: Here, too, I'm missing a formal significance test, e.g. via ANOVA.
Fig. 4 was included to show that "location with low soil water storage capacity tended to have large lambda values". The kruskal wallis test (used because of outliers and non-normal distribution) shows that the mean TI values of the storage classes are significantly different. We will include that information in the manuscript.

References

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70 (1968): 213-220

Gupta, V.K., and S. Sorooshian. 1985. The relationship between data and the precision of parameter estimates of hydrologic models. Journal of Hydrology 81:57–77.

Hahn, C., V. Prasuhn, C. Stamm, P. Lazzarotto, M.W.H. Evangelou, and R. Schulin. 2013. Prediction of dissolved reactive phosphorus losses from small agricultural catchments: calibration and validation of a parsimonious model. Hydrological and Earth System Sciences Discussions 10:1465-1510.

Lane SN, Brookes CJ, Heathwaite AL, Reaney S. (2006). Surveillant science: challenges for the management of rural environments emerging from the new generation diffuse pollution models. Journal of Agricultural Economics, 57(2), 239-257.

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

Milledge DG, Lane SN, Heathwaite AL, Reaney SM. (2012). A Monte Carlo approach to the inverse problem of diffuse pollution risk in agricultural catchments. Science of the Total Environment, 433, 434-449.

Rankinen, K., T. Karvonen, and D. Butterfield. 2006. An application of the GLUE methodology for estimating the parameters of the INCA-N model. Science of The Total Ennvironment 365 123–139.

Reaney SM, Lane SN, Heathwaite AL, Dugdale L. (2011). Risk-based modelling of diffuse land use impacts upon instream ecology. Ecological Modelling, 222, 1016–1029.

Winsemius, H.C., B. Schaefli, A. Montanari, and H.H.G. Savenije. 2009. On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. Water Resources Research 45:W12422.

Yapo, P.O., H.V. Gupta, and S. Sorooshian. 1996. Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. Journal of Hydrology 181:23–48.

Calculation of weighted kappa
(J. Cohen, 1968)

Stägbach

RRP yearly average DRP load (rows) **versus** SCIMAP locational risk (columns)

	1	**2**	**3**	**4**
1	8658	1962	30	0
2	3	222	44	24
3	0	11	4	1
4	0	0	7	0

$weights
[1] "squared"
$kappa
[1] 0.2584333

RRP DRP load at highest runoff **versus** SCIMAP locational risk

	1	**2**	**3**	**4**
1	8659	1459	8	0
2	2	736	44	0
3	0	0	32	18
4	0	0	1	7

$weights
[1] "squared"
$kappa
[1] 0.5438046

RRP yearly average DRP load **versus** SCIMAP global locational risk

	1	**2**	**3**	**4**
1	8744	1878	28	0
2	0	180	87	26
3	0	5	10	1
4	0	0	7	0

$weights
[1] "squared"
$kappa
[1] 0.2927247

RRP DRP load at highest runoff **versus** SCIMAP global locational risk

Groups	**1**	**2**	**3**	**4**
1	8742	1377	7	0
2	2	686	94	0
3	0	0	30	20
4	0	0	1	7

$weights
[1] "squared"
$kappa
[1] 0.5639102

Lippenrütibach

RRP yearly average DRP load (rows) **versus** SCIMAP locational risk (columns)

Groups	**1**	**2**	**3**	**4**
1	3512	355	68	3
2	0	18	41	2
3	0	0	7	3
4	0	0	0	1

$weights

[1] "squared"

$kappa

[1] 0.2950730

RRP DRP load at highest runoff **versus** SCIMAP locational risk

	1	**2**	**3**	**4**
1	3454	215	20	0
2	58	128	40	6
3	0	30	45	2
4	0	0	11	1

$weights

[1] "squared"

$kappa

[1] 0.6789908

RRP yearly average DRP load **versus** SCIMAP global locational risk

	1	**2**	**3**	**4**
1	3692	236	10	0
2	0	39	22	0
3	0	2	5	3
4	0	0	0	1

$weights

[1] "squared"

$kappa

[1] 0.454832

RRP DRP load at highest runoff **versus** SCIMAP global locational risk

Groups	**1**	**2**	**3**	**4**
1	3558	130	1	0
2	130	71	29	2
3	4	67	4	2
4	0	9	3	0

$weights

[1] "squared"

$kappa

[1] 0.5884168

Table with wkappa values

	Stägbach			Lippenrütibach	
	SCIMAP locational risk	SCIMAP global locational risk		SCIMAP locational risk	SCIMAP global locational risk
Stägbach					
RRP – average DRP load	0.26	0.29			
RRP – DRP load at highest runoff event	0.54	0.56			
Lippenrütibach					
RRP – average DRP load				0.30	0.45
RRP – DRP load at highest runoff event				0.68	0.59