The manuscript “Development of a conceptual model of the hydrologic response of tropical Andean micro-catchments in Southern Ecuador” by Crespo et al. aims at better understanding the flow systems at the study sites by model hypotheses testing. Although this could be a potentially very interesting paper, it is not yet quite developed to the point, which is rather surprising given the amount of senior expertise among the authors. While the approach of comparing different catchments with distinct responses is very valuable, the methods used and the analysis of the results remain very superficial.

My main concern is that while the authors promise hypotheses testing, all they do is comparing merely two model hypotheses. In the model calibration and evaluation section they then address possible shortcomings of the model, which however and UNFORTUNATELY remain very speculative instead of actually *testing* these hypotheses. Thus, including more extensive and rigorous hypotheses testing (e.g. Clark et al., 2011) could make this paper a very important contribution for better understanding headwater catchment behaviour in the Andes.

Slightly changing the actual research question/objective from “developing a working model” to “improving our understanding of the processes driving the hydrologic response” and thus the resulting models could give the authors a better feel for what really might be interesting in this study: it is not developing more or less working models for 4 basically irrelevant catchments, but it is rather the wider scope of getting a better understanding of the underlying processes in this region, i.e. use the models as a learning tool (cf. Clark et al., 2008; Fenicia et al., 2008).

My second major concern is the model itself, which seems not to be very consistent, as well as the calibration strategy and results which are only insufficiently documented. For example no mention is made about parameter uncertainty.

Furthermore, I have the feeling that the manuscript is too long and has to be significantly shortened as it gives much information that is irrelevant for the paper and/or redundant. I would also strongly encourage the authors to have the manuscript proofread by a native English speaker.

If the authors carefully addressed the detailed comments below, I would be happy to see the paper eventually published as it could be very interesting to wide parts of the community. However, quite some work still needs to be invested to bring the manuscript to actual publication standard.

Detailed comments:

p.2476, l.7 and elsewhere: I would suggest the use of less “strong” words such as “support” and “adequate” instead of “confirm” and “correct”

p.2476, l.11-14: sentence does not make much sense. Rephrase or omit.

p.2476, l.22: why in particular in South America? Please omit.

p.2477, l.8: wide parts of the hydrology community, including myself, might not be familiar with paramo ecosystems. Please explain.

p.2477, l.11 and elsewhere: it is stated that stream flow is sustained by lateral matrix flow. Do you mean “sustained” such as in low flow or do rather mean “dominated”? If lateral matrix flow is actually sustaining low flows, I would be good to document this in a bit more detail as I feel it is a quite uncommon process. Can you please add a conceptual sketch of a typical hillslope? This would help clarifying things. As I understand it, you have relatively shallow (impermeable?) bedrock and water runs off on the sloped C-horizon/bedrock interface, driven by the elevation head, rather than by the pressure head (as it would be for groundwater). Is this correct? However, there is one thing I cannot quite follow: the Ksat values given in table 2 seem generally not to show significant differences between near-surface and deeper horizons. Being quite high, these Ksat values imply fast draining of the respective horizons – what is then sustaining stream flow during prolonged dry periods?

p.2477, l.20-23: can be omitted as same information is in the preceding paragraph

p.2477, l.27 and elsewhere: should read as “C-horizon”

p.2477, l.26-29: this is rather trivial. Can be omitted.

p.2478, l.1-5: this also only repeats information already given above. Please omit.

p.2478, l.4-5: please be specific: what is comprised in the top horizons?

p.2478, l.8-15: not really surprising. Can be omitted.

p.2478, l.25: “most adequately” is quite a stretch here as only two possible model structures were tested.

p.2479-2482: although good catchment descriptions are necessary for the reader to get a better feel for the study environment, the description of the study sites here is too long and redundant. I would encourage to authors to considerable shorten it to 1 page maximum, as much of the information is not actually relevant to the paper (e.g. pH values) and where relevant given in the references provided.

p.2479/Figure 1: Please provide more meaningful maps. Instead of 2 context maps, just use one context map with the locations of the catchments and a close-up map of the catchments with elevation and the locations of climate stations and gauges.

p.2480, l.15 and elsewhere: Although the authors obviously did considerable work in this area, the number of self-citations is a bit excessive. It would be very beneficial to the paper if the authors diversified the references and reduced the number of self-citations, which by the way also make the paper more credible and more likely to be read by others.

p.2482: although soil horizon depths are given, the depth of the regolith (C-horizon), arguably the most important for sustaining base flows is not given. Are there no estimates available from literature?

p.2482, l.19: please give more details on the nature of this intra-day curve!

p.2482, l.21: how did you define "acceptable"?

p.2483, l.2: Please give a reference for the area weighted elevation method. Is there actually a significant elevation gradient present in the region? If so, please give a value and a reference.

p.2483, l.7-8: can be omitted.

p.2483, l.9-10: what kind of control measurements? Did they support the methods applied otherwise?

p.2483/Figure 2: please show interception reservoirs in figure 2!

p.2484, l.7: why was canopy interception fixed? Why do not use it as calibration parameter?

p.2484, l.7-8: is the interception loss fraction constant over time? Wouldn't we suspect that it changes with wetness conditions? How did you determine it? Based on land-cover?

p.2484, Eq.5-7: why do you show a simple linear reservoir concept with threshold in such a complicated way. Just call it what it is.

p.2485, l.6-7: What is direct overland flow? SOF is commonly standing for Saturation Overland Flow. Besides that, saturation excess is NOT generated when the rainfall intensity is higher than infiltration rates – this is rather infiltration excess overlanf flow or Hortonian overland flow (HOF).

p.2485, l.8-10: not sure if I understand this correctly. The amount of water in S1 exceeding S1max is multiplied by the runoff coefficient to produce SOF? If so, what is happening to the remaining water exceeding S1max? Please clarify! Please also specify the runoff coefficient. Is it the long term average runoff coefficient? How did you calculate it?

p.2485/Figure 2: there are couple of concerns I do have about the model structure. (a) what is the reason to include the thresholds TS2 and TS3 in the model? They do not have any effect as once the reservoir level drops below TS2 or TS2, the water remains there. This is different to S1 where water below TS1 can be transpired by plants, thus effectively introducing a threshold. TS2 and TS3 therefore can be removed. (b) why is TLS1>TS1 and TLS2>TS2? Wouldn't we expect an earlier onset of vertical percolation (in particular given the high hydraulic conductivities) than lateral flow? Thus, TLS1 and TLS2 need to be smaller than TS1 and TS2. (c) what happens if S2>S2max and S3>S3max? where is the excess water stored? What happens to the percolation in such cases? Please clarify!

p.2485, l.14-18: Please show the reader a table with the parameters, specifying if they are free calibration parameters, fixed as a result from values in literature or as a result from observations. Please also give the parameter ranges.

p.2485, l.24-25: please specify “surface interception”. Is it the same as litter interception?

p.2486, l.1-10: the calibration strategy needs to be described in much more detail. The readers are neither shown the parameter ranges for sampling, nor are they told about assumptions with respect to prior distributions, the use of likelihood measures and the number of MC realizations. With 14 free parameters, the model would require quite a large number of MC realizations (probably >10^6). Further the reader is, unfortunately, not shown any information on parameter uncertainty (e.g. 5/95% ranges or dotty plots). Besides that, and in the light of the problems reported in correctly modeling low flows, it would be worth investigating if the use of a master recession curve (e.g. Lamb and Beven, 1997) could help to pre-identify the storage coefficients. Another questions that arises is, why do the authors not consider multi-objective and multi-criteria calibration, for example using the Nash-Sutcliffe efficiency of the log of the flows and/or calibrating the model also to the flow duration curves? This would strongly increase confidence in the realism and thus the predictive capabilities of the models.

p.2486, l.13-25: it would be interesting to see how the runoff coefficients vary between dry and wet seasons!

p.2487, l.4-7: irrelevant here. Can be omitted.

p.2487, l.10-11: sentence not clear. What do you mean by water regulation capacity?

p.2487, l.11-12: can be omitted.

p.2487, l.14ff: how different are calibration and validation periods in terms of climate? If they are very similar, the validation results should be expected to be close to the calibration. The real challenge is getting the validation right in periods that are considerably different to the calibration period.

p.2488-2491: rather than giving a lengthy, flag waving description of what the model is able to do it would MUCH more instructive to focus on what the models fail to do! Thus, much of page 2488 can be condensed.

p.2488, l.12 and elsewhere: this is not the residence time! What you mean is the “time scale” or 1/storage coefficient

p.2488, l.15: it can be due to underestimation of precipitation but it is not limited to that.

p.2488, l.20: what are concave, saturated plateaus?

p.2488, l.21-23: is it really wider or is it only wider on the log-scale?

p.2489, l.10-12: evaporation has obviously to be estimated with elevation corrected temperature, whose environmental lapse rates are arguably much more stable than the precipitation elevation gradients the authors used in their area weighted elevation method. I would thus encourage the authors to actually *test* the effect of that!

p.2489, l.12-14: why not test a non-linear set-up then??

p.2489-2491: there is plenty of speculation in the discussion of the results and this leaves ample room for really interesting improvements. I would strongly encourage the authors to actually test at least some of the hypotheses they discuss here, e.g. can the storage coefficient of the base flow be fixed (see above), does adjusting evaporation to elevation help (see above), make interception capacities free calibration parameters, use test non-linear structures (see above), spatially distributed interception and soil moisture stores (e.g. Fenicia et al., 2008), use distributions rather then step functions to characterize threshold behaviour (e.g. Clark et al., 2008).

p.2492, l.15: please specify "groundwater". How is base flow sustained? That has got to be some sort of saturated flow.

p.2493, l.2-4: it would be good to link this to findings of other process studies. Thus please include the two references Penna et al., 2011 and Hrachowitz et al., 2011.

p.2493, l.8-10, Figure 5: please show that information on plots normalized by the total runoff

Table 2: what is the relevance of pH and SOM here? Please omit.

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H.V., Wagener, T. and Hay, L.E. (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research 44, W00B02

Clark, M.P., Kavetski, D. and Fenicia, F. (2011), Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research 47, W09301

Fenicia, F., Savenije, H.H.G., Matgen, P. and Pfister, L. (2008), Understanding catchment behaviour through stepwise model concept improvement. Water Resources Research 44, W01402

Hrachowitz, M., Bohte, R., Mul, M.L., Bogaard, T.A., Savenije, H.H.G. and Uhlenbrook, S. (2011), On the value of combined event runoff and tracer analysis to improve understanding of catchment functioning in a data-scarce semi-arid area. Hydrology and Earth System Sciences 15, 2007-2024

Lamb, R. and Beven, K. (1997), Using interactive recession curve analysis to specify a general catchment storage model. Hydrology and Earth System Sciences 1, 101-113

Penna, D., Tromp-van Meerveld, H.J., Gobbi, A., Borga, M. and Dalla Fontana, G. (2011), The influence of soil moisture on threshold runoff generation processes in an alpine headwater catchment. Hydrology and Earth System Sciences 15, 689-702