Reviewer 1

1. Does the paper address relevant scientific questions within the scope of HESS? YES

2. Does the paper present novel concepts, ideas, tools, or data? YES, however, the need for WALRUS or other new specific lumped conceptual approaches is not necessarily clear when approaches like SUPERFLEX are available.
In the companion GMDD-paper, in which we present WALRUS, we explain the need for a rainfall-runoff model for lowland catchments. We wonder if a flexible approach, such as SUPERFLEX, can deal with feedbacks between reservoirs, especially between surface water and groundwater. In addition, we prefer a fixed over flexible model structure because it is much easier to apply by practitioners.

3. Are substantial conclusions reached? Perhaps not substantial, but the conclusions are adequate.

4. Are the scientific methods and assumptions valid and clearly outlined? YES

5. Are the results sufficient to support the interpretations and conclusions? YES

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? YES

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution? YES

8. Does the title clearly reflect the contents of the paper? YES

9. Does the abstract provide a concise and complete summary? YES

10. Is the overall presentation well structured and clear? YES

11. Is the language fluent and precise? YES

12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? YES

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated? The number of Figures could easily be reduced to only capture the results that support the main conclusions of the paper.
Although the paper contains many Figures, we think that they are necessary to fully describe out results. Note that reviewer 2 says: "This is a very comprehensive paper testing a newly developed model for areas with low topographic relief. The investigation is extensive, including sensitivity analysis, uncertainty analysis, extreme conditions analysis, and investigation of multiple modelpredicted variables and routines."

14. Are the number and quality of references appropriate? YES

15. Is the amount and quality of supplementary material appropriate? YES

The manuscript by Brauer et al. (2014) puts a new rainfall-runoff model WALRUS to the test in the Hupsel Brook catchment and Cabauw Polder. Overall the paper is well written and easy to follow. The paper is clear and the objectives for assessing the WALRUS model and the conclusions drawn are made clear. Therefore I recommend that the paper be accepted for publication after some minor revisions. Below some comments on the paper and some minor errors are provided.

1. In the calibration section 3, the parameters are referred to as having physical connotations, which is questionable. The fact the parameter names have a sense of some physical connotation is merely a measure of convenience.
We agree with the reviewer that physical interpretations of parameters of conceptual models should be handled with care. To stress that, we added "(although we stress that physical interpretations of parameters of conceptual models should be handled with care)" to Section 3.2 (Section 3.1 in HESSD). The parameters were given names based on their function in the model. For example, the quickflow reservoir constant determines the storage-outflow relation of the reservoir determining the fast runoff response. We added "(the effect of each parameter will be investigated in Sect. 5.1)" to show that we checked whether the real effect of the parameters is the same as we intended.

2. How were the ranges selected for the model parameters specified on Pg 2103 Ln 11-12?
The ranges were selected based on exploratory runs.

3. It would be useful to see the Nash-Sutcliffe efficiencies for ET, dv and dg and to discuss these aspects of the model performance.
We discussed the model performance for ET_{act}, d_{V} and d_{G} by comparing the lines in the Figures. We chose not to compute the goodness-of-fit essentially because catchment effective values cannot be compared to point measurements, especially for d_{V} and d_{G}.We added the following sentence to Section 4.1 (this is a new Section called "Validation Methods"): "Observations of groundwater depth and storage deficit were used for a qualitative appreciation of the internal model dynamics."

4. In Sections 4.3 and 4.4 the calibrated Hupsel Brook catchment model is tested on two extremes. Given that WALRUS is a lumped conceptual model, its application to events outside of the range of calibration events seems reason enough that issues arise with the model capturing the extremes.
It is true that most models, lumped as well as distributed,

have difficulties simulating extremes. The spatial variability in threshold exceedance is an extra challenge for lumped models. For wet situations, the wetness index in WALRUS captures part of the spatial variability in the degree of saturation and consequently in active flowroutes. For dry situations, the nonlinear relation between groundwater depth and surface water level on the one hand and groundwater drainage on the other hand captures the effect of decreasing conductivity between groundwater and surface water when headwaters run dry. We performed the tests reported in the HESSD-paper to investigate if these approaches to simulate the effect of spatial variation on runoff production in a parsimonious way also yield good results in extreme situations.

5. On Pg 2109 Ln 27-30. This last sentence is not clear. Also, relating point measurements to a catchment effective model parameter has no clear purpose. This sentence should be rewritten or removed.
We added "groundwater and storage deficit" to clarify what we mean with "variables".

Minor remarks:

Pg 2093 Ln 1: Replace specially with especially
Done.

Pg 2096 Ln 14: Avoid the use of impossible and use extremely difficult or some less definitive variant.
Done.

Pg 2103 Ln 2: Insert the after as.
Done.

Pg 2113 Ln14: Replace in with is in cV in inversely proportional.
Done.

Pg 2117 Ln 4: Neverteless spelt incorrectly.
Corrected.

In Figure 6 and 7, labels should be given to identify the two catchments in the left and right hand side plots.
Somehow the labels got lost in the process of optimizing the Figures. We added labels top Figures 5, 6 and 7 and specified the caption of Figure 9.

Reviewer 2

This is a very comprehensive paper testing a newly developed model for areas with low topographic relief. The investigation is extensive, including sensitivity analysis, uncertainty analysis, extreme conditions analysis, and investigation of multiple modelpredicted variables and routines.

While the majority of the paper is wonderfully written and explained as is, I have two concerns:

-Figure 4 implies to me that some of your parameter values are not identifiable: The best value as identified by HydroPSO also occurs for some of the parameters in parts of the parameter space that are very different from other parts of the parameter space where NS values also appear to be high (e.g. cw and cg for Cabauw polder). It may be worth investigating if a different parameter set with the same level of fit reproduces the time series in a very different way. Another option see if you arrive at the same values by starting HydroPSO at a few different initial parameter sets. Singleobjective optimization algorithms are often sensitive to this. You do a very nice job later in the paper investigating the impacts of parameter sensitivity and uncertainty, but I think its worth investigating just how robust your optimized parameter set is, given that this is a focus of a significant part of the manuscript.
We agree that parameter estimation remains difficult and the outcome of a particle swarm optimization algorithm is always different from a Monte Carlo analysis. If we use the best (in terms of Nash-Sutcliffe efficiency) parameter set from the Monte Carlo analysis, we obtain similar plots for the validation runs, with similar Nash-Sutcliffe efficiencies. This indicates that, even with only four parameters, there is some risk of equifinality.

-Section 5.1 (Parameter identifiability): Performing two different sensitivity analyses is comprehensive, and your figures that display these analyses are very nice. However, changing a single parameter value at a time or investigating first order effects does not address the most important issue when it comes to equifinality how much do the parameters interact? If the model is computationally inexpensive to run, Id suggest applying a simple sensitivity analysis, e.g. Method of Morris, which measures the amount of interaction per parameter.
This is certainly an important aspect. However, exploring this in detail is outside the scope of this paper. Besides investigating the interactions between two parameters (which we touched upon with the response surfaces in Figure 13), it would also be good to look at interactions between three or four parameters and the nonlinearity in the interactions between the parameters because the role of the parameters changes in time (see Fig. 11).

Minor suggestions follow:

Page 2096, line 26: remove e.g.
Done.

Page 2097, lines 8-9: a little awkwardly phrased!
We changed the sentence to "In the Hupsel Brook catchment, many hydrological variables have been measured intermittently since the 1960s."

Page 2097, lines 22 23: all were measured or observations

consider revising, reads awkwardly.
We changed it to: "net radiation was measured and the sensible and ground heat fluxes were estimated from wind and temperature profiles."

Page 2098, line 8: Id remove considered as a catchment in this study unless this distinction is important, or replace considered with treated?
We changed "considered" to "used".

Page 2101, lines 24-29: run-on sentence
We structured the sentence more: "The influence of water management in the Cabauw polder is discernible in three ways: (1) discharges remain high in summer due to surface water supply, (2) on 6 May 2008 discharge suddenly dropped to zero as a result of the increase of weir elevation, and (3) on 16 November 2007 and 15 October 2008, discharge increased because the weir was lowered."

Page 2103, line 17: parameters should not be plural
Done.

Page 2105, line 9: time should be plural
Done.

Figure 14 difficulty distinguishing different dashes maybe use a dot-dash combination instead?
We changed the lines to dots.

Reviewer 3

General Comments: Overall, the paper entitled The Wageningen Lowland Runoff Simulator (WALRUS): application to the Hupsel Brook catchment and Cabauw polder offers a nice comparison between two very different catchments in a similar climatic regime. The Cabauw polder experiences heavy influence of tile draining to ensure efficient runoff of water when the groundwater table is shallow. The WALRUS model shows flexibility to account for an influx of water into the catchment area from sources other than precipitation and the interaction and feedback that accounts, in part, for artificial drainage networks. Overall the work appears to be well done, although I found the structure of the paper to be an inefficient way to portray the good work that was completed.

In general, I would suggest 1) to separate out the methods applied in this study to its own section for an easy to see overview of your methodologies
The set-up of this paper is not standard, but in our eyes logical. We did not include a general Methods Section in the revised version, but we added Section 3.1 "Calibration Methods" and Section 3.2 "Validation Methods". In these Sections the description of the methods for each Section can be found. We also adapted the last paragraph of the introduction.
and 2) include additional discussion and references comparing the results of this work to others.
Comparing results to other studies is always a valuable contribution, but we think that a detailed discussion on this point is outside the scope of this paper. We do intend to perform a model comparison study to compare the performance of WALRUS to other rainfall-runoff models. Is the surface water-groundwater interaction enough to model artificial drainage networks? Had this been something traditionally absent in studies trying to model catchments with artificial drainage, which found poor results with overly simplistic models?
In the companion paper for GMD, we give examples of results reported in similar catchments with different models. In the introduction of the GMDD-paper, we write: "Examples of the resulting problems are presented by Bormann and Elfert (2010), who used WaSiMETH (Schulla and Jasper, 2007) and Koch et al. (2013), who used SWAT, both in north-eastern Germany."
References:
Bormann, H. and Elfert, S.: Application of WaSiM-ETH model to Northern German lowland catchments: model performance in relation to catchment characteristics and sensitivity to land use change, Adv. Geosci., 27, 110, 2010.
and:
Koch, S., Bauwe, A., and Lennartz, B.: Application of the SWAT Model for a tile-drained lowland catchment in north-eastern Germany on subbasin scale, Water Resour. Manag., 27, 791805, 2013..
What did we learn in this study that was not known in past experience running this model? The quality of the research is good, however the structure of the paper should be refined. It is for this reason that I recommend major revision.

Specific Comments:

Optimized parameter sets seemed not to be very behavioral. Meaning, the performance of the model was not sensitive to most model parameter values. This suggests interaction between parameters in the model, which hints at the problem of equifinality. This is briefly discussed, but I think it would be constructive to include a bit more discussion. It might be to answer a question such as Could you constrain the model in the future to help alleviate this problem?
This is an important discussion subject and a topic we are currently working on. For example, we are investigating methods for parameter regularisation to constrain parameters during calibration. We think it would be outside the scope of this paper to go into this topic in detail.
Why do we see the differences in parameter sensitivity depending on the objective function? This is a result that I would expect to see, however it would be good to offer a

bit more discussion as to why you think this might be the case.
We added the sentence "As expected, the parameter sensitivity changes with the objective function, which indicates that the importance of a parameter changes between high and low flows."

Page 2097 – Line 16-17 Before 1988 the method of Thom and Oliver (1977) has been used and since 1989 the method of Makkink (1957). This sentence is a bit awkward. Please revise to make your point clearer.
We spit the sentence: Before 1988 the method of Thom and Oliver (1977) has been used. Since 1989 the method of Makkink (1957) has been used.

Page 2103 Line 14 – Many dates are shown as time periods that have necessary time series data for calibration. However, it is not clear to me what time periods were actually used for calibration.
The periods used for calibration are given in Section 3 on Calibration: "For the calibration, we used hourly data of the periods November 2011–October 2012 (Hupsel) and October 2007–September 2008 (Cabauw).".
Was there a warmup period to initialize states within the model?
We added the following sentences to Section 3.1: "It was not necessary to use a warming-up period. The initial groundwater depth for the calibration period was calibrated together with the parameters. The other initial states followed from the observed discharge at the start of the period, the stage-discharges relation and the model equations and parameters (see Brauer et al., 2014)."
Was the time period used for calibration similar to that used in validation?
The main validation runs had a length of one year as well. Shorter periods were used for the flood, drought and management case studies. We summarized all periods in a (new) Table in the (new) Methods Subsection of the Validation Section. If not, what might be the consequences of this (would a longer validation period cause a degradation in performance of the model over time, or would the variations between the time series used for validation be averaged out over time?)?
A validation run over tens of years (not shown) does not show degradation in model performance over time.

Page 2115 Line 23 – "are not physical" should read something like "are not physically feasible."
Changed.

Table 1 – ET is listed on the table, however it is not very clear if this is ETpotential or ETactual. Please make notation regarding ET consistent with the rest of the paper.
We changed it to ET_{pot}.

Additional remarks

We found an error ourselves on page 2103, lines 22–28, in the explanation of the lower value of c_{W} for the Cabauw polder. We changed this Section:

"When comparing the Cabauw polder to the Hupsel Brook catchment, differences in parameter values can be observed and explained (although we stress that physical interpretations of parameters of conceptual models should be handled with care). Parameters c_{V}, c_{G} and c_{Q} are higher, indicating that all flow is slower. Parameter c_{W} is smaller, causing ~~earlier~~ later activation of quick flowroutes (at lower storage deficits). Compared to the Hupsel Brook catchment, the clayey soil in the Cabauw polder is less permeable, leading to slower groundwater flow (c_{G}) and a slower response of groundwater to changes in the unsaturated zone (c_{V}). There are more cracks, gullies and drainpipes per unit area ~~(c_{W})~~, but quickflow is activated later (c_{W}) because connectivity is limited. Quickflow is slower (c_{Q}) because slopes of land surface (overland flow) and drainpipes are more gentle. It is not a coincidence that the drainage density increases when permeability decreases. Farmers install drainpipes and dig gullies when ponding hampers agricultural activities, animals (moles, mice and muskrats) dig more burrows to drain their dens and cracks occur more quickly in clayey soils."