Review of the manuscript „Results from a full coupling of the HIRHAM regional climate model and the MIKE SHE hydrological model fora Danish catchment" by Larsen et al., Hydrol. Earth Syst. Sci. Discuss., 11, 3005–3047, 2014

This manuscript is studying the influence of one-way and two-way coupling on the performance of the HIRHAM and Mike SHE models. Particular emphasis is put on the influence of the data transfer interval (DTI) using two-way coupling. The results are validated by observations from the Skjern river catchment. The effort presented in this manuscript is very interesting for the weather and climate research communities because more accurate and consistent simulations of the water cycle in catchments are needed for many end users and decision makers.

Based on my comments below, I recommend major revision before the manuscript can be published. This is mainly related to the clarification of the methodology and a more detailed interpretation of the results.

Major points:

1) Coupling issues: The authors are distinguishing between one-way and two-way coupling but do not explain the difference between both approaches in a sufficient manner. I suggest that the authors add a dedicated section to introduce and discuss the different methods:

- In one way coupling, the hydrological model is only driven by the output of the regional climate model. What is the difference in horizontal grid increments between both models in this case? It would be useful to apply a very high resolution version of Mike SHE in this case but this is not mentioned in the manuscript, or at least, I did not find it. In any case, it can be expected that the output of Mike SHE is strongly biases by the inconsistency of model physics in the regional model and resulting errors in the forcing data. However, the authors claim that better results should be expected due to a separate calibration of the regional climate and the hydrological models. This is not convincing because significant biases occur in the HIRHAM model output.
- In two-way coupling, data exchange is taking place between the surface layer of HIRHAM and the soil-vegetation layer of Mike SHE. In this case, the authors mention correctly that in the inner domain the interaction of model physics is different that the domain outside of the catchment. The influence of this effect should be discussed more in detail, as this may propagated significantly in the inner domain consisting only of a few grid boxes. Particularly, I am missing a more detailed explanation of the physics used in the data exchange. The HIRHAM model output is applied for driving Mike SHE and the surface energy balance is fed back to HIRHAM. Thus, in both models, the fluxes in the surface layer are influenced by surface layer stability namely Monin-Obokhuv stability theory (MOST). This needs to be calculated in HIRHAM in order to extrapolate the surface layer variables to the lowest atmospheric layer. However, nothing is mentioned to ensure consistency here.

Please add a paragraph explaining more in detail the data exchange demonstrating that the physics between HIRHAM and Mike SHE is fully consistent.

2) DTI: It is a very interesting issue to study the influence of the DTI. However, most of the performance can be proposed not only by the coupled simulations but also by some physical considerations. For instance, if the DTI becomes too large, model imbalance issues can be expected because the evolution of model variables does not consider the change of fluxes in the meantime. Was this effect observed? A time scale of 10-30 min seems to be appropriate under unstable conditions to track the change of fluxes, which is supported by the results whereas shorter time periods are likely not necessary. I would appreciate a more extensive discussion of these considerations. Particularly, the time scale where model inconsistencies can be expected should be taken into account in the choice of the DTI.

3) Model grid increments: I am missing a discussion of model resolution issues. Using a grid increment of 11 km of a hydrostatic model, significant precipitation biases can be expected. Model output is also quite coarse for driving a hydrological model. Why did the authors not downscale the regional model results in the catchment for providing better forcing data and realizing a more realistic two-way coupling? It can be expected that the model results will improve at grid increments of approximately 1-3 km because land-surface heterogeneity are better resolved and the parameterization of convection can be avoided. What is the effect of different resolutions of Mike SHE on the performance of two-way coupled simulations?

4) The authors did not convince me that at the present stage two-way coupling should show a worse performance than one-way coupling. Even if HIRHAM and Mike SHE were not calibrated together – here I am wondering what "calibration" means for a regional climate model – the consistent modeling of water fluxes should lead to better results, if the same MOST is taking into account. Otherwise, the degradation of the results may be explained by this effect?

Minor points:

Introduction: Please consider also Kunstmann and Stadler, J. Hydrology 2005, and Smiatek et al. Env. Mod. Software 2012 as well as Shresta et al. Mon. Wea. Rev. 2014 as examples of coupled modeling.

P. 3016: Why did the authors not perform HUV runs with perturbed model physics? In this case, model uncertainties can be assessed in a more realistic manner.

P. 3017: It does not make much sense to distinguish between the different domains 1-5 because they deviate just by a few pixels. Please note that the real resolution of a model corresponds to 3-4 times the grid increment so that difference between the results will hardly be significant. I suggest just concentrating on the catchment.

P. 3018: It is more common to use V instead as WS for horizontal wind.

P. 3023, l. 9-10: I do not understand this sentence. Please clarify.

Fig.3: It is nice to see that a reduction of DTI results in a decrease of rmse. It is obvious that this effect is reduced dealing with large-scale variables such as Rg and Ps. However, please explain why the rmse in D5 is often much smaller. This is a strange effect.

Fig. 5: Please add the observations to these figures. It is not clear and not sufficiently discussed in the text why the rms is larger for the TI runs rather than for HUV and CV. Maybe this increased sensitivity is realistic, as the coupled run allows for a more accurate simulation of LSA feedback?

Fig.6: The gray lines can hardly be distinguished. I suggest using different colors. I do not understand why the authors conclude a better performance of HUV. TI seems to be at least similar for short CTIs and 60-min CTI (CV) are likely biased by the strong delay of the update. I think this figure supports the value of two-way coupling of the models. Moreover, the differences between the domains are likely not significant and may be due to different sampling errors in the observations.

Fig.7: Same as in Fig.6. The coupled runs seem to simulate a more reasonable variability. This should be supported by the observations. Why are these missing on this figure? It is very important to add these in a thick black line for example.

Fig.8: It is quite optimistic to compare a grid-box value with a point measurement when considering fluxes. The sub-grid scale variability of land-use in the grid box is certainly quite variable. It only makes sense to show these comparisons, if further downscaling of the model resolution is applied. The good performance of Q is another promising aspect of coupled modeling.

Fig.9: Please add the observations, too.

In general, the Figs. support the value of coupled simulations, as pointed out in the discussion and the conclusions. However, I am missing some additional aspects (see also the major issues above). There is a great potential to improve the simulations by increasing model resolution. Why is this not considered? Most of the errors of the simulation of precip are due to incorrect model physics (e.g. cloud microphysics, convection parameterization) and most of the variability in the perturbations may be explained by this effect. What can be done in the future to reduce this? Otherwise, coupled runs will be of limited value. I would appreciate a more extended discussion of these tradeoffs.

Grammar:

- P. 3016, l. 22: Insert comma before “a varying level …”
- P. 3018, l. 24: Remove period after “ … °C”