Response to reviewer # 1

This study presents the results from a coupled regional climate model HIRHAM with MIKE SHE hydrological model, which also includes SWET land surface model. The coupling between the models are only done over the Skjern catchment , which is an interesting feature of this new tool. With this new tool, a series of real data numerical experiments with coupled and uncoupled models are presented to explore the influence of coupling frequency and internal variability of the atmospheric model. The study is quite interesting, and the multiple simulations along with comparisons to observations are comprehensive. However, at the same time, there are several shortcomings in the current version of the paper also, which needs to be addressed before being suitable for publication.

In general, the manuscript is poorly written with lot of grammatical mistakes and not well organized.

We will have an English native speaker to help improving the English language and the organization of the manuscript.

The authors conclude that the coupled simulations give poor results because the coupled model is not tuned or calibrated. However, there are no results presented in the current manuscript that supports their conclusion.

I understand the comment as a call for; 1) proof that the coupled setup provides poorer results than the uncoupled HIRHAM due to the lack of combined calibration and *not* 2) a call for proof that the coupled setup is actually poorer than the uncoupled and the reply here is written accordingly. Also because the proof for the latter (2) is evident in figures 3, 5 and 8.

The question of the need for a coupled calibration is centred on either providing the right answer for the wrong reason or providing the wrong answer for the right reason. As stated, both models are refined, tuned or calibrated (phrasing depends on community; climate/hydrology) to reproduce observations and any change in forcing data (surface scheme and climate input data), constituting significant elements of each model, are likely to worsen the results over the area in question.

One way of indirectly supporting this statement would be to actually show these forcings to have a significant impact on the model outcome. In the present paper this is especially seen in figure 5 where the RMSE values are both higher and include more variability for coupled runs. Another figure showing the influence of model forcing is seen in Butts et al. (submitted) where distributed evapotranspiration output for a one-week period is seen for 1) MIKE SHE forced with observations, 2) MIKE SHE forced with HIRHAM input (one-way with no feedback), 3) MIKE SHE forced with HIRHAM input (two-way including feedback) and 4) for HIRHAM alone. From this figure the influence of the coupling is evident as i), MIKE SHE produces higher evapotranspiration with observation input as compared to using HIRHAM input and, ii) the feedback between models significantly influences the coupled setup outputs as the two-way coupled evapotranspiration is higher than one-way coupled (no feedback from MIKE SHE).

We are a bit cautious on expanding the overall volume of the paper, but in the revised version we will elaborate on the issue of dynamics and coupled calibration.

A way of directly supporting the statement of the need for coupled calibration would be to include just that; coupled calibration. However, this is beyond the reach of the present study. Also, the question of coupled calibration is used in the present paper as a general comment for the coupling of systems calibrated and tuned individually to provide physical sense on their own in terms of energy- and water balance closure (providing an answer – right or wrong – for the right reason). We could add a short reference to coupled ocean-atmosphere studies where the experience is very similar.

It could also be potentially influenced by their limited area coupling. In their approach, they couple the two models over a small catchment only, which is less than 0.1% of the total atmospheric domain. So, does it create a very different patch of land surface over Skjern catchment compared to the adjacent cells, where HIRHAM uses its own land surface model? If there are significant differences in soil temperature due to difference in partitioning of surface energy fluxes by the new model, it can generate local circulations, which can influence the simulated variables. This needs to be discussed.

In spite of the coupling only covering 0.02% of the RCM domain our results show that it nonetheless has significant impacts on the land surface variables. This can e.g. be seen from figure 5, where the cells outside the coupled domain (domain) show much less difference between the RCM (TUV) and the coupled model (CV) as compared to the cells within the Skjern catchment (e.g. domain 1)" This is an interesting result in itself and documents that the local partitioning of surface energy fluxes within the Skjern catchment has a significant impact on the land surface variables and that those are not only determined by boundary effects (advection). We will include a discussion on the issue of local circulations deriving from the now area-limited forcing of the coupled area.

1. Pg. 3007, Line 10: The authors here can definitely not say that it is due to the calibration of the models alone. There could be many other reasons. This needs to be addressed. Again, we are cautious on expanding the paper volume but we would be happy to rephrase here into something along the lines of: "It is discussed whether this may be caused by the lack of calibration of the coupled model" in both abstract and discussion.

2. Pg. 3007, Line 18: Change "ranges" to "spread". Ok.

3. Pg. 3009: The introduction is poorly written, the authors discuss about future global climate predictions and importance of uncertainty in climate models, and then at the end, present what they do in this study. A more focused discussion on the scientific questions they want to answer with this new tool and the motivation behind this work, would strengthen this section. We agree that this section sets of very broadly, something which is done to frame the study into the context of highlighting the need for further knowledge and investigation of the atmosphere – land surface process interrelations. We however agree to shorten the broader perspectives and more clearly emphasize the specific motivation for the study.

4. Pg.3009, Line 6-16: These works were mostly related to short term simulations using mesoscale atmospheric models, not climate simulations. The spatio-temporal time scales of these studies compared to the preceding introduction are different. This needs to be rephrased. We agree and will rephrase.

5. Pg. 3009, Line 26-29: For example, read York et al. 2002, Jiang et al .(2009), Anyah et al. (2008). C926

These are some very interesting papers. Thank you. Since York et al. (2002) uses a single atmospheric model grid, Jiang et al. (2009) uses a simple non-3D SIMGM groundwater model and Anyah et al. (2008) (as described in Miguez-Macho et al. 2007) uses Darcy's law for the horizontal groundwater fluxes we believe the statement still holds true whereas we would be happy to modify from:

"To our knowledge, no studies have been reported on long term simulations (more than a few days) with couplings between a regional climate model and a 3-D groundwater–surface water hydrological model using catchments larger than a single regional climate model grid point."

To:

"To our knowledge, no studies have been reported on long term simulations (more than a few days) with couplings between a <u>distributed</u> regional climate model and a <u>full</u> 3-D groundwater–surface water hydrological model using catchments larger than a single regional climate model grid point."

Also, we will add these suggested papers in the introduction literature review.

6. Pg. 3011, Line 19: Clarify "the undercatch corrected precipitation". Good suggestion.

7. Pg. 3012, Line 9: Does it mean that the fluxes measured over forested area was used for agricultural site? Explain the rationale behind it. Missing data alone does not justify this approach. This approach is well documented and approved in peer reviewed papers for these specific data (Ringgaard et al 2011). We however clearly acknowledge the pitfalls of this approach and will make a short comment on this in the paper.

8. Pg. 3013, Line 20: Is this calibration done for every particular year? The calibration is described in another paper (submitted – as described in the reference) where sensitivity analysis and inverse modelling is applied for a one-year period with subsequent validation for another period.

9. Pg. 3015, Line 11: Clarify "safety regulations". So, the coupling is based on reading and writing of output files? Yes.

10. Pg. 3015-3017: The description of the different simulations requires a Table with two sub-cateogries: "coupled and uncoupled simulations" followed by experiment name and description. Otherwise, it becomes too difficult for the readers to follow which experiment is which, and it is very annoying. The evaluations are performed in terms of RMSE and MAE. The results could be interpreted better by using MSE and examining the contribution of model bias, variance of the simulated variables and the correlations. And, the use of Taylor diagrams would be even more appropriate to present the results for comparison of different variables with multiple simulations, coupled or uncoupled. Adding a table for a general simulation overview is a good idea that we could

try to fit in without much added volume. Regarding performance measures we have used MAE as this is more intuitive to understand the actual absolute differences and the choice between RMSE and MSE is simply a matter of reducing the plotted numbers for a good and balanced overview. Again, due to volume concerns we are hesitant to add Taylor diagrams as an additional figure as we believe the present figures still provide important information used in the discussion regarding model performance, data transfer frequency and variability.

11. Pg. 3020 - 3021: See above comments. Answered above.

12. Pg. 3023 and Pg. 3030, Line 10: This needs to be rephrased. See above comments. We can broaden this statement based on our answer above.

13. All figures have a very small font size which is not readable. Good observation – we will revise.

References

- Butts, M., Drews, M., Larsen, M.A.D., Lerer, S., Rasmussen, S.H., Groos, J., Overgaard, J., Refsgaard, J.C., Christensen, O.B. and Christensen, J.H. (submitted): Embedding complex hydrology in the regional climate system – dynamic coupling across different modelling domains.
- Larsen, M. A. D., P. Thejll, J. H. Christensen, J. C. Refsgaard, and K. H. Jensen (2013a) On the role of domain size and resolution in the simulations with the HIRHAM region climate model, Clim. Dynam., 40, 2903–2918, doi:10.1007/s00382-012-1513-y.
- Maxwell, R.M., Lundquist, J.K., Mirocha, J.D., Smith, S.G., Woordward, C.S. and Tompson, A.F.B. (2011) Development of a Coupled Groundwater–Atmosphere Model. Mon Weather Rev, Vol 139, 96-116, doi:10.1175/2010MWR3392.1.

Response to reviewer # 2

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:

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: Good suggestion. The other reviewer suggested a table for an improved simulation overview: We will add this and discuss the differences between simulations a bit further in the text- 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? In each and every simulation MIKE SHE uses a 500 m resolution and HIRHAM a 11 km resolution. Also, in every case of data transfer the data is interpolated and aggregated into the grid of the coarsest model (11 km). 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. Do you mean transferring data to HIRHAM from MIKE SHE at a finer resolution than 11 km? HIRHAM does not include the possibility to use land surface information on a finer grid scale than the overall resolution. 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. What we state is that each of the two model systems are calibrated or tuned individually and that the coupled setup provides a significant change in forcing data dynamics and levels. This discussion is very general and related to the overall motivation of the study. We agree that HIRHAM, being a regional climate model, has a certain degree of both coarseness and bias transferred into MIKE SHE. However, the intent of the study is a first attempt to demonstrate the performance of the coupled system to locally provide more detail in the land surface input to HIRHAM; spatially, temporally and in absolute levels. For the issue on bias correction see below.

- 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.
  It is entirely correct that there is a probable physics inconsistency in the current prototype version of the HIRHAM-MIKE SHE coupled model, and we completely agree that this could better be highlighted in the manuscript. We will add a paragraph on this issue. The noted inconsistency will be addressed comprehensively in future versions of the coupling. As mentioned above in the present state of the model the surface energy balance of MIKE SHE effectively overwrites that of HIRHAM thus changing the energy balance calculated through its simple land surface scheme in a very abrupt manner. Effects relating to this inconsistency are likely to be seen most strongly along the boundary of the inner model domain, and probably accounts for a prominent part of the observed differences between coupled and uncoupled simulations. That said, the errors we observe are relatively small. One reason for this is probably due to the scale difference and thus to cell averaging and cancellation of errors, when feeding the MIKE SHE surface back to HIRHAM. Also, HIRHAM is itself a model code comprised of an atmospheric mesoscale model coupled to an "external" land surface scheme, which here is replaced by MIKE SHE. Another factor to consider is the data transfer frequency, which defines how often the surface forcing from MIKE SHE is updated. Unsurprisingly, we see improvements at higher data transfer frequencies corresponding to more dynamic interactions between land surface and atmosphere.

1) 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. Exactly. Was this effect observed? The decrease in RMSE with a more frequent data transfer is indeed an indicator for this yes. 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. As shown the DTI definitely matters on the coupled model performance for four out of six variables and optimally the DTI should reflect the temporal resolution of the most dynamic of these exchange variables. A varying exchange depending on the atmospheric stability is an interesting perspective to optimize both model performance and computation time. We will add a smaller discussion on the temporal scale (dynamics) of variables in relation to the DTI and model performance.

2)

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. Yes we are approaching the limits of the model in terms of resolution as also documented in Larsen et al. (2013). Other papers such as Roosmalen et al. (2010) also addressed the issues of HIRHAM bias although here related to SST's. 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? We do not agree that 11 km grid data is too coarse for driving a hydrological model. The Danish national water resources model, from which the present hydrological model has been cut out is successfully forced by 10 km grid data (Stisen et al., 2012). We did not downscale the HIRHAM data for two reasons: i) we do not have sufficient detailed local observations to downscale to 500 m grid used by the hydrological model; and ii) we wanted to preserve the energy and water balances of HIRHAM as part of the experiement. 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? Yes, we can hope so, but we do not know as for the current setup as this is not done for the fully coupled setup and also, it depends on which performance variable is investigated as well as the temporal and spatial scale. Both spatial and temporal biases will e.g. affect the water available for evapotranspiration. In Larsen et al. (2013) the performance of HIRHAM is investigated showing a tendency for the larger 11 km resolution domains to better reproduce seasonal precipitation and temperature as compared to smaller 5.5 km resolution domains. This study was used to assess the domain to be used in the coupled setup and possibly indicates that the resolution threshold for HIRHAM, being hydrostatic, is within this range.

1) **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? To address these issues separately:

- For the two-way coupled simulations the MIKE SHE feedback is sent back to HIRHAM as opposed to one-way where HIRHAM uses its own land surface scheme. The comparison between one-way and two-way is therefore relevant for MIKE SHE only as one-way in this sense essentially means uncoupled (for HIRHAM). Therefore; yes the uncoupled HIRHAM output performance is worse than the coupled as shown in figures 3 and 5. Regarding one-way versus two-way coupling MIKE SHE results, the performance is shown in figure 8 and here uncoupled results (observation data forcing) are generally better than one-way coupled results (HIRHAM data forcing). Exceptions however include; agricultural LE and soil heat fluxes (G) for all sites (!). Reviewer #1 suggested a table for simulation overview. We have added this and hope that it will improve the overview of the simulations performed.

- In the hydrological community calibration refers to adjustment of parameter values (coefficients in equations) to make the model match observational data better. In the climate modelling community the term calibration is rarely used. Instead the term refinement or tuning (e.g. of precipitation scheme parameters or of albedo) is often used. So what we suggest is that climate models could be subject to systematic calibration/refinement/tuning as also suggested by Bellprat et al. (2012)
- As described above the same MOST is not taken into account for results being compared.

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. Good suggestions. We will include these.

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. HUV runs are indeed eight perturbed uncoupled simulations. The added table that we suggest will provide an improved simulation overview in a new table as described.

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. We agree that the uncoupled HIRHAM results will hardly be affected by these domain differences but we use these to distinguish between the model outcome with varying degrees of coupling – something which is clearly visible in figure 3.

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

P. 3023, l. 9-10: I do not understand this sentence. Please clarify. As stated above through precipitation parameterization (hydrostatic scheme) and energy balance tunings (albedo). If urged, we will add this.

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. Due to DOM5 having 0% coupling (catchment/domain overlap).

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? The point about an expected increase in sensitivity and variability for the TI runs is interesting. However, I am not sure we fully understand the issue on adding observations to figure 5. The Y-axis is an RMSE level (for simulations against observations and for the entire period). Do you mean adding the period mean observations to a secondary Y-axis? I am not sure this would add any meaning to the figure. Regarding adding observations to figure 7 and 9, see the response for

figure 7.

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. Indeed, for precipitation, there is a need to distinguish between dynamics and cumulative performance and for the results shown in this figure we agree that the coupled performance is not necessarily worse than uncoupled. We will add this statement to the discussion and change the colors of the TI runs. Good points.

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.

We have chosen not to add the observations to figure 5, 7 and 9. The figure will be (even) more incomprehensible and voluminous and we would have to discuss the specific dynamics of observations against simulations (for the shown period only) which would make an already extensive paper even larger. This is not the intent of the figure. The intent is instead to show that the CV, HUV and TI fall in distinct groups of comparable patterns related to their simulation condition (being coupled (perturbed), coupled (varying DTI) or uncoupled (perturbed)). A discussion on hour-to-hour simulation performance against observations would also be limited by the issue of comparing an 11 km grid output with gridded observation data from point stations (is this really feasible?) and this would have to be addressed again increasing the paper volume. Instead we address observations against simulations by longer term run statistics.

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. We definitely agree, whereas this was the only choice since we did not include downscaling (as explained above). We will add a comment on this issue.

Fig.9: Please add the observations, too. See the answer above.

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? Due to the hydrostatic nature of HIRHAM, unfortunately. 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. We will further discuss these issues regarding perspectives on RCM resolution (hydrostatic versus non-hydrostatic), larger catchment (reduced edge effect), DTI, etc.

Grammar:

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

References

- Bellprat, O., S. Kotlarski, D. Lüthi, and C. Schär (2012), Objective calibration of regional climate models, J. Geophys. Res., 117, D23115, doi:10.1029/2012JD018262.
- Larsen, M. A. D., P. Thejll, J. H. Christensen, J. C. Refsgaard, and K. H. Jensen (2013a) On the role of domain size and resolution in the simulations with the HIRHAM region climate model, Clim. Dynam., 40, 2903–2918, doi:10.1007/s00382-012-1513-y.
- van Roosmalen, L. V., Christensen, J. H., Butts, M. B., Jensen, K. H., and Refsgaard, J. C.: An intercomparison of regional climate model data for hydrological impact studies in Denmark, J. Hydrol., 380, 406–419, doi:10.1016/j.jhydrol.2009.11.014, 2010.
- Stisen, S., Højberg, A. L., Troldborg, L., Refsgaard, J. C., Christensen, B. S. B., Olsen, M., and Henriksen, H. J.: On the importance of appropriate precipitation gauge catch correction for hydrological modelling at mid to high latitudes, Hydrol. Earth Syst. Sci., 16, 4157–4176, doi:10.5194/hess-16-4157-2012, 2012.