We appreciate the reviews from Referee#1 and believe that the comments and suggestions have significantly improved our manuscript. In the following, we address specific reviewer comments.

Anonymous Referee #1

General comments

1.This paper focuses on the description and evaluation of a procedure for reducing the time of model spin-up, which is a commonly adopted strategy to initialize integrated/ coupled hydrological models such as ParFlow.CLM. In my opinion, there is a core issue within this paper related to its very basic idea, i.e., the assumption that an equilibrium state (achieved over no matter how many years of forcing data) can represent a correct (or even reasonable) initial catchment state. Although I acknowledge that this is a common assumption, I believe that not only it is not true in general, but the number of cases where this could be reasonable is limited, in theory, only to catchments where i) the land use do not change over time and ii), most importantly, the inter-annual variability of the weather forcing is very small. The latter point is equivalent to the assumption that a single year (or two, three) of forcing data can be considered representative of the whole climatic regime of the catchment, an hypothesis that is never realistic in practice. Unfortunately for hydrologists, catchments are always dynamic systems and never in a state of equilibrium; therefore, I am afraid that the whole procedure proposed in the paper is not worth the effort from the very beginning. Instead, the only way to achieve a correct or reasonable initial state is to use a "warm-up" procedure, where the model must be run using a long enough time series of forcing data before the period of interest; the necessary warm-up duration will be obviously catchment-specific and can be evaluated by starting the model with two or more different initial guesses and checking that after the warm-up all the simulations converged to the same final (dynamic) state.

We agree with the reviewer comments regarding the short comings of equilibrium based initialization for initializing coupled/integrated hydrologic models. Despite these shortcomings this initialization method is commonly used. This technical note provides a method for improving the efficiency of this commonly used initialization technique. While in the land surface modelling community various experiments have been performed across multiple sites and models to assess the impact of initialization approaches and spin-up criteria (Yang et al., 1995, Rodell et al., 2005) on simulated response, the issue of model initialization has not been fully explored for the coupled or integrated hydrologic models. As we stated in our objectives, here the goal was to reduce the spin-up period for equilibrium based initializations. The equilibrium based initializations have been used previously for exploring land surface-groundwater coupling (Kollet and Maxwell, 2008) and assessing the impact of climate change on groundwater-land surface interactions using an integrated hydrologic model (Ferguson and Maxwell, 2010). Here, similar to the Project for Intercomparison of Land Surface Parameterization Schemes (PILPS), only one year of forcing data is used for ParFlow.CLM spin-up. The adjustment method introduced can be examined when multiple years of forcing data is used for the warm-up period. We could not examine the impact of multiple years of forcing on the spin-up time due to the intensive computational demand. In Ajami et al. (2014), the impact of different initializations approaches as examined in the literature was briefly discussed and we refer the reader to that summary. The issue of initialization is very important particularly in coupled or integrated hydrologic models and coordinated efforts to perform such experiments across multiple models and sites are required in hydrologic modelling community.

The following point is added to the revised Introduction:
The two most common initialization approaches in coupled or integrated distributed hydrologic models are: (1) initial depth to water table is specified at a certain uniform depth below the land surface (Kollet and Maxwell, 2008) and the impact of initialization is reduced through recursive simulations over either a single or multiple years of forcing data, until equilibrium conditions are reached, which are usually related to spin-up criteria based on changes in groundwater heads (Refsgaard, 1997) or changes in water and energy balances (Kollet and Maxwell, 2008); or (2) the model is initialized from a fully saturated condition and simulations are continued until modelled baseflow matches the observations (Jones et al., 2008). Equilibrium based initializations have been utilized previously for exploring land surface-groundwater coupling (Kollet and Maxwell, 2008) and assessing the impact of climate change on groundwater-land surface interactions using an integrated hydrologic model (Ferguson and Maxwell, 2010).

Further, we added application of equilibrium based initialization in the objective section of the revised manuscript:

The objective of the current study is to develop a *hybrid* spin-up approach that significantly reduces the number of years of spin-up required for model state equilibrium. The equilibrium based initialization represents a correct initial state for catchments in which the land use does not change over time and the inter-annual variability of atmospheric forcing is very small: assumption that are common to most simulation frameworks. This technical note provides a method for improving the efficiency of this commonly used initialization technique. The performance of the proposed approach in reducing the spin-up period for a catchment scale application of the ParFlow.CLM model is evaluated against the standard continuous recursive simulation approach that is commonly applied for land surface model spin-up, and referred to here as the baseline spin-up approach.

2. Another issue of this paper regards the lack of important details, such as (at least) a brief description of ParFlow.CLM, and some steps of the procedure that are not described with sufficient clarity. See below in the list of specific comments.

A brief description of ParFlow.CLM is added to the revised manuscript. Please see Section 2.1. Section 2.2 is revised to include further details about the approach.

Specific comments:
1. Page 6971, line 20: please define "service unit".
Definition of a service unit is added in the revised manuscript.
(a service unit is equivalent to 1 hour of time used by one processor)

2. Page 6972, section 2.1: despite the title, no description whatsoever of the model is provided, but only a description of the two catchments.

A brief description of ParFlow.CLM is added at the beginning of section 2.1 as follows:
ParFlow is a 3D variably saturated groundwater flow model that solves the mixed form of the three-dimensional Richards equation for the subsurface (Ashby and Falgout, 1996; Jones and Woodward, 2001; Maxwell et al., 2014). ParFlow has a fully integrated overland flow simulator (Kollet and Maxwell, 2006) and performs routing of the ponded water on the land surface via the kinematic wave equation. The Common Land Model (CLM 3.0) (Dai *et al.*,

2003) is integrated into ParFlow to simulate water and energy fluxes at the land surface (Maxwell and Miller, 2005; Kollet and Maxwell, 2008). ParFlow.CLM versions 605 and 653 were used for the Skjern River and Baldry simulations respectively, which are described below. The terrain following grid of Maxwell (2013) is not implemented in these modelling set-ups.

3. Page 6975: lines 7-19: this section is rather difficult to follow. Is the DTWT function used to re-initialize the model spatially variable or uniform? And the resulting DTWT distribution after re-initialization? Also, it is not clear how the "best performing" DTWT functions were chosen: what objective function was used to evaluate the best performance, root mean square difference, mean absolute error or the semi-variogram?

The DTWT function produces spatially distributed DTWT for re-initialization. The procedure for generating spatially distributed DTWT is as follows: 1) develop the DTWT function based on percent changes in mean annual DTWT values across the domain for six cycles of ParFlow.CLM simulations (global DTWT function), and 2) Implement the DTWT function at every grid cell to re-initialize the ParFlow.CLM model. The updated DTWT at every grid cell depends on its initial value. To clarify this point, the manuscript is revised as follows:

The empirical DTWT functions calculated above estimate percentage changes in mean annual DTWT as a function of simulation year. To predict spatially distributed mean annual DTWT from a global DTWT function, the mean annual DTWT from the final cycle of the ParFlow.CLM spin-up simulation for every grid cell is used as the initial value to successively estimate DTWT distributions as a function of simulation year. These DTWT distributions are based on the predicted percent change values from the global DTWT function.

Here we are comparing the performance of multiple DTWT functions against the baseline simulation using multiple objective functions. As presented in our results a single objective function does not constantly perform best for all the cases and each of the objective functions provides a summary statistic regarding a certain aspect of the model performance. While mean absolute error (MAE) and root mean square difference (RMSD) provide an overall average of model error, RMSD is more sensitive to extreme values. Percent bias gives information about the average tendency of the model prediction to be larger or smaller than the baseline simulation. We revised the manuscript as follows for the Skjern River sub-catchment section:

Optimum parameter values for single and double exponential DTWT functions were obtained using nonlinear least squares method. Performance of the single and double exponential DTWT functions in predicting 14 years of DTWT were compared against ParFlow.CLM baseline spin-up simulations (years 7 through 20) of Ajami et al. (2014) to find the optimum empirical DTWT function for the Skjern River sub-catchment. Post-simulation analysis indicates that global DTWT functions based on domain or catchment averaged percentage change values are better predictors of DTWT response compared to local DTWT functions developed for every grid cell. Instability of local DTWT functions occurs in grid cells where percent changes in DTWT oscillate between positive and negative values through initial spin-up simulations. Spatial distribution of these grid cells are shown in Fig S1.

Calculated RMSD and percent bias relative to the baseline spin-up simulations indicate that global double exponential functions using ParFlow.CLM spin-up simulations 2 to 6 provide a better fit compared to various single exponential functions obtained from different spin-up simulation years (e.g. 2 to 3, 2 to 4, etc.). Because the first six cycles of ParFlow.CLM simulations were the same between the baseline spin-up simulations and DTWT distributions from DTWT functions presented in Fig 4, comparisons were made with simulations 7 to 20 of the baseline spin-up approach of Ajami et al. (2014).

As can be seen from Fig. 4a, the mean annual DTWT over the domain derived from the single exponential functions (fitted to percentage change data from simulations 2 to 6) under-predict the baseline spin-up simulations, due to their consistent small underestimates in comparison to double exponential functions fitted to the same data points. Only for the mean absolute error (MAE) calculated at each pixel do single exponential functions based on simulations 2 to 6 perform slightly better and produce smaller errors on average than the double exponential functions (Fig. 4b). It should be noted that the percent bias in mean annual DTWT for simulation cycle 20 is -1.6% for the domain based double exponential function and -6.2% for the single exponential function, with both functions derived from simulation cycles 2 to 6. Therefore, single exponential functions are not further examined in re-initializations of the DTWT. In terms of mean DTWT across the domain (Fig. 4a), the catchment delineated double exponential DTWT function provides a better prediction and the smallest mean bias when compared to the function based on the entire model domain. However, Figure 4b indicates that the mean absolute error values are slightly smaller for the domain based double exponential function. The higher MAE of the catchment based double exponential function is a result of slightly more regions with over and underestimated DTWT values that contribute to a good overall mean DTWT (Fig. 4a), but contains more errors spatially compared to the domain based double exponential function.

The section is followed by an overall conclusions in the revised manuscript:
In summary, double exponential functions are chosen as they have less bias compared to single exponential functions and there is very little difference in terms of MAE amongst predictions. The choice is further supported by the RMSD and semi-variograms.

4. Page 6976, lines 2-3: why is the pressure head profile in the UZ adjusted with a (basically) instantaneous distribution, taken from the last day of the sixth cycle, while the re-initialized DTWT is assumed as an annual mean? I see a possible lack of consistency that should be discussed.

To clarify this point, the manuscript is revised as follows:

Section 2.2.
Sensitivity of spin-up functions across multiple criteria and variables showed that the estimated spin-up period based on mean annual DTWT were more stable when compared to other spin-up criteria, such as changes in the mean DTWT for the last day of recursive simulations (Ajami et al., 2014)….

In the adjusted pressure head approach, the hydrostatic equilibrium assumption is used in regions between the new DTWT and the initial DTWT. The ParFlow.CLM pressure head distribution is adjusted to begin at the new pressure head from the initial WT such that the vertical profile is maintained (Fig. 3). This adjustment may represent a lack of consistency in the proposed approach as the DTWT function estimates mean annual DTWT, while pressure head adjustments in the unsaturated zone are taken from the last day of the sixth cycle of ParFlow.CLM. While it is possible to use DTWT values from the last day of simulations to develop a DTWT function, estimated DTWT values from such a function exhibit larger variability and result in a larger bias. For the Skjern River sub-catchment, percent bias between the estimated DTWT values from the DTWT functions and the baseline simulation of Ajami et al. (2014) were -4% and -1.6% for the DTWT functions based on the last day and mean annual DTWT values respectively.

5. Page 6977, lines 1-17 and Fig. 3: I am quite puzzled by these results. From Fig. 3a, I would expect that i) the MAE of the Exp2-Catchment curve decreased with time, not the contrary, especially after year 14, and ii) the MAE of the two Exp1 curves was larger, not smaller, than the Exp2 curves. Have the authors any explanation for this?

Here the MAE is calculated on a pixel basis by computing the average absolute difference between the estimated DTWT from the DTWT functions and the baseline spin-up simulation. We included the formulas for calculating the objective functions in the revised manuscript to clarify this point. Figure 4a shows the mean annual DTWT over the domain and it indicates that the Exp2-Catchment function results in the smallest mean bias and the single exponential functions have the worst overall mean. Figure 4b shows the MAE calculated at each pixel and it indicates that the Exp2-Catchment function has slightly more regions with over and underestimated DTWT values resulting in a good overall mean, but contain more errors spatially. Meanwhile, the single exponential function estimates result in consistent small underestimates which produce slightly smaller errors when averaged spatially but has a worse overall mean.

Therefore, i) Figure 4b of the revised manuscript indicates that the performance of DTWT functions deteriorate for later time period (simulations 7 through 20) and the MAE increases in later simulations. ii) Similarly, average model error on a pixel basis (MAE) is smaller for the two single exponential functions than the Exp2 curves as shown in Figure 4b.

To clarify this point, the manuscript is revised. Please see the response to specific comment #3.

6. Page 6977, lines 18-29: it is not clear how the semi-variograms were calculated. Was the mean annual DTWT used?

Mean annual DTWT at every grid cell was used to calculate semi-variograms. The manuscript is revised to describe the procedure.

To investigate this result further, three empirical semi-variograms were generated. As the impact of an east-west spatial trend in the mean annual DTWT values was evident in the semi-variograms, the trend should first be removed from the mean annual DTWT values. To remove the trend, a plane was fitted to the observed mean annual DTWT values, with an equation of the form:

$$z = ax + by + c \qquad (6)$$

where *a*, *b* and *c* are fitted coefficients, x and y are the coordinates of every grid cell, and z is the mean annual DTWT. Residuals are computed by subtracting the estimated mean annual DTWT from Equation 6 from the observed mean annual DTWT values. Finally, the semi-variogram of the residuals as a function of distance is calculated.

7.Page 6978, lines 20-22: this sentence is not clear, please rephrase.
The sentences are rephrased in the revised manuscript:

While in both re-initializations, DTWT and subsequently groundwater storage volume were the same at the start of the simulations, unsaturated zone storage of the hydrostatic equilibrium option was drier than the adjusted pressure head option. Additional ParFlow.CLM simulations after re-initialization ensured equilibrium of groundwater storage. As can be seen from Fig. 6a, hydrostatic re-initialization results in a deeper WT at equilibrium (simulation 12) relative to the baseline equilibrium year (simulation 20). Higher DTWT values of the hydrostatic option at equilibrium correspond to smaller groundwater storage and subsequently larger unsaturated zone storage compared to the baseline spin-up (Fig. 5). It should be noted that in ParFlow.CLM, groundwater and unsaturated zone storages are not explicitly determined by fixed size compartments and the extent of an unsaturated zone is determined by the location of the water table. Percent changes in mean annual unsaturated zone storage between the last two years of recursive simulations were 0.1% for the hydrostatic equilibrium and 0.3% for the adjusted pressure head re-initializations, indicating unsaturated zone equilibrium at different threshold levels.

8. Page 6979, lines 5-9 and Fig. 6: from the figure I cannot see how the adjusted vertical pressure distribution produces better results than the hydrostatic profiles, nor I can see the bias with the latter. Perhaps, would be a good idea to show the experimental pdf (articula) of the differences along with their spatial distribution.

Figure 6a shows that the difference between DTWT distributions from the hydrostatic equilibrium option and the baseline simulation is mostly positive, while for the adjusted pressure head option the differences in DTWT values are negative in the upper part of the catchment. We generated the kernel density plots of the differences; however, the figure was not informative especially when the bandwidth was set the same for both density plots. Therefore, we only included the spatial distribution of the differences in the revised manuscript and changed the colour scheme of the Figure to present this bias.

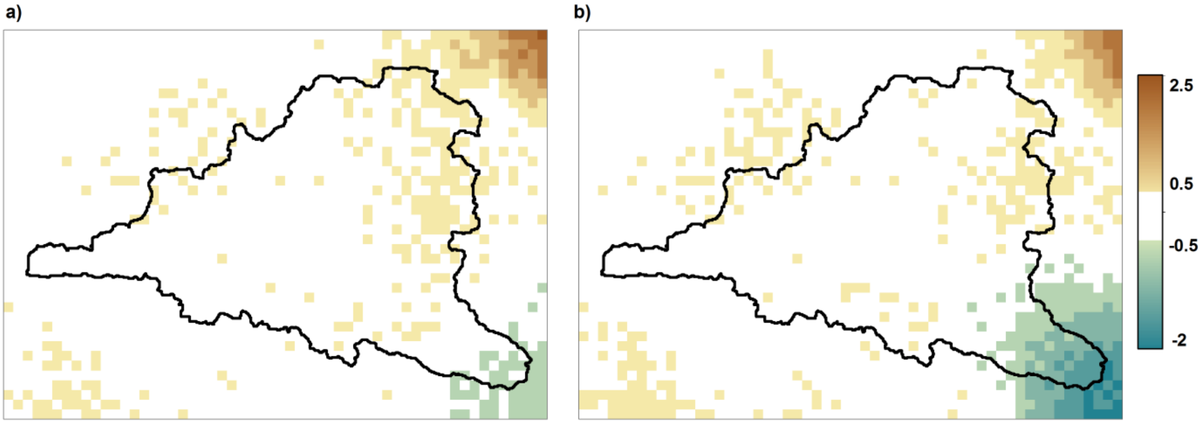

Figure 6. Differences in equilibrium DTWT between ParFlow.CLM simulations after re-initializations and ParFlow.CLM after 20 years of baseline spin-up simulations in (m), where a) is based on hydrostatic pressure distribution above the water table for the initial condition, while b) is based on adjusted pressure head distribution above the water table for the Skjern River sub-catchment. White regions correspond to gird cells where the differences in equilibrium DTWT are less than 0.5.

9. Page 6979, line 16: why does the smaller Baldry catchment require more service units than the larger Skjern catchment? Is it because the former has a larger grid size due to a better DEM resolution?

A major factor in the Baldry catchment requiring more service units is that it required 40% longer simulations to reach equilibrium. The total number of service units for one year of simulation is larger for the Baldry sub-catchment than the Skjern River sub-catchment despite its lower number of computational nodes (467712 nodes in Baldry compared to 909440 in the Skjern River sub-catchment). It is difficult to solely attribute the increases in computational time to the DEM cell size. Based on our experience, increases in the number of vertical nodes result in longer computational time. Here, the increases in computational time in Baldry are due to multiple factors. As indicated in the revised manuscript, two different versions of ParFlow have been used for these catchments and we used different options for storing the CLM output files (silo versus PFB). For the Skjern River sub-catchment, CLM output files were saved as distributed silo files and after every model restart (every 15 days), a one processor job was submitted to un-distribute the silo files and save them as PFB files. For the Baldry sub-catchment, the PFB option was used and the un-distribution of PFB files was performed with the main ParFlow TCL script that uses 64 processors. Therefore, this set-up has led to unrealistic increases in service units for the Baldry. Based on the ParFlow user forum, it seems that the issue with the un-distributed PFB files can be resolved by using different setting when compiling the ParFlow.CLM code.

We should note that the number of processors for every catchment scale simulation was determined by performing parallel efficiency tests.

10. Page 6981, lines 1-2: I do not agree that the proposed procedure "has the potential to assist in parameter calibration". Due to equifinality, if a wrong initial state is used, such as the one likely to achieve by assuming equilibrium, a calibration procedure could lead to strongly biased parameters.

We acknowledge the reviewer concerns here. We are not attempting to address equifinality here, we simply present a technique that can make previously used calibration approaches more efficient. We have removed this statement in the revised manuscript. It should be noted that a period of spin-up has been often implemented during calibration. We revised the Summary section as follows:

Previous efforts in calibrating coupled or integrated hydrologic models required a spin-up process after every parameter update (Stisen et al., 2011; Weill et al., 2013). Development of a computationally efficient spin-up approach will enable this type of systematic calibration of integrated or coupled hydrologic models.

11. Page 6990, Fig. 5: I am surprised that the hydrostatic equilibrium procedure underestimates the groundwater storage even from the start of the simulation. If the DTWT at re-initialization is the same as for the adjusted pressure profile, how can the Authors explain that large bias?

To address reviewer comment, the manuscript is modified as follows:

While in both re-initializations, DTWT and subsequently groundwater storage volume were the same at the start of the simulations, unsaturated zone storage of the hydrostatic equilibrium option was drier than the adjusted pressure head option. Additional ParFlow.CLM simulations after re-initialization ensured equilibrium of groundwater storage. As can be seen from Fig. 6a, hydrostatic re-initialization results in a deeper WT at equilibrium (simulation 12) relative to the baseline equilibrium year (simulation 20). Higher DTWT values of the hydrostatic option at equilibrium correspond to smaller groundwater storage and subsequently larger unsaturated zone storage compared to the baseline spin-up (Fig. 5). It should be noted that in ParFlow.CLM, groundwater and unsaturated zone storages are not explicitly determined by fixed size compartments and the extent of an unsaturated zone is determined by the location of the water table. Percent changes in mean annual unsaturated zone storage between the last two years of recursive simulations were 0.1% for the hydrostatic equilibrium and 0.3% for the adjusted pressure head re-initializations, indicating unsaturated zone equilibrium at different threshold levels.

Figure 5b shows time series of groundwater storage for the equilibrium year for three cases, baseline simulation, adjusted pressure head and hydrostatic options. The equilibrium year corresponds to simulation cycles of 10, 12 and 20 for the adjusted pressure head, hydrostatic equilibrium, and baseline simulations, respectively. Figure 5b caption is revised to clarify this:

Figure 5. Comparison of a) unsaturated and b) groundwater storages of ParFlow.CLM equilibrium year using the hybrid and baseline spin-up approaches (Ajami et al., 2014). The equilibrium year corresponds to simulation cycles of 10, 12 and 20 for the adjusted pressure head, hydrostatic equilibrium, and baseline simulations, respectively. The dynamics of groundwater and unsaturated zone storages are closely reproduced by the adjusted pressure head distribution approach relative to the baseline spin-up approach for the Skjern River sub-catchment.

12. Page 6992, Fig. 7: there seems to be a spatial pattern, with streaks of DTWT overestimation in the south of the catchment. How can this be explained?

The contours of DTWT overestimation occurred along the direction of flow lines from high elevation areas in the catchment toward the catchment outlet. As indicated in the manuscript, the performance of global DTWT functions were deteriorated in high elevation areas. The Figure caption is revised as follows:

Figure 7. Differences in equilibrium DTWT of Baldry ParFlow.CLM simulations after re-initialization with the adjusted pressure head distribution above the water table and ParFlow.CLM after 28 years of baseline spin-up simulations in (m). The contours of DTWT overestimation are along the direction of flow lines from high elevation areas toward the catchment outlet.

Technical corrections

1. Page 6970, lines 18-19: change the sentence to "The issue of model initialization is important for hydrologic predictions as the initial state has a major impact on the catchment's model response".

It is revised.

2. Page 6979, line 21: correct "particular".

It is corrected.

3. Page 6974, line 17: DTWT was 3 m only for the Skjern catchment.

Initial DTWT for the Baldry sub-catchment is included in the revised manuscript.

4. Page 6977, line 20: change "semi-variance" to "semi-variogram".

It is modified.

References:

Rodell, M., P. R. Houser, A. A. Berg, and J. S. Famiglietti (2005), Evaluation of 10 methods for initializing a land surface model, *J. Hydrometeorol.*, 6(2), 146-155.

Yang, Z. L., R. E. Dickinson, A. Henderson-Sellers, and A. J. Pitman (1995), Preliminary study of spin-up processes in land surface models with the first stage data of Project for Intercomparison of Land Surface Parameterization Schemes Phase 1(a), *J. Geophys. Res.*, 100(D8), 16553-16578.