Response to Reviewer #1 comments

The authors manually calibrate three land surface models (CLM, Noah, VIC) by using observed monthly streamflow at streamflow gauges for one period (_~5 years) and validate for the independent other period (~5 years). The three models were run from 1951 to 2015 to produce root zone (~60 cm) soil moisture products. The authors use these soil moisture products to analyze Indian agricultural drought events including severity, frequency, and drought extent. The results found that there is larger uncertainty in crop growing season than the monsoon season. The large uncertainty is mainly due to the difference in model parameterizations – different soil moisture persistence. The results suggest using multi-model ensemble for Indian drought monitoring. For model setup and calibration, model evaluation, and analysis of differences in model parameterizations, the paper shows some major deficiencies in its general appearance. Therefore, I recommend a major revision of the manuscript.

We thank the reviewer for his/her insightful comments. We have made every possible effort to address the reviewer's comments in an adequate manner (in below).

1. Model setup:

(a) Spin-up period: Is a spin-up period run? If no, please explain reason. If so, how long is run for each model for this spin-up period? Was soil moisture equilibrium state including deep soil layer checked?

Yes, we run a spin-up period for all models to avoid (undesirable) influence of initial conditions. The spin-up period was set to 1951-2015. For this period, we ran each model and generated an initial state file using which the simulations were conducted for the entire period. Moreover, we performed an exploratory analysis to make sure that each model is in the stable condition (equilibrium state) from the beginning simulations.

(b) I do not think that you can use daily meteorological forcing data to run Noah and CLM? In general, hourly surface forcing data are used to drive such land surface models. How to divide daily meteorological forcing data into hourly time scale?

The variables other than daily precipitation, maximum and minimum temperatures, and wind speed were generated using the VIC model, which uses the MTCLIM method. As we have reported in our manuscript, the effectiveness of the MTCLIM algorithm has been evaluated using the observations from the flux-tower for various ecosystems of the world (Bohn et al. 2013). The Noah and CLM models internally disaggregate daily forcing to sub-daily time steps. The output from these models was again aggregated to daily time step for the analysis.
We would like to mention here that the method chosen for the disaggregation of daily values to hourly ones would not have any (substantial) effects on the resulting monthly soil moisture simulations, which is eventually used for the drought analysis.

(c) It is not clear how to calibrate Noah model. Why are depth of soil layers, Zilintikevich coefficient, surface runoff parameter and bare soil evaporation component selected? Is any sensitivity test performed or does the selection just depend on your own experience? Which are surface runoff parameter and soil evaporation component? What possible values do you use? How to manually tweak these values for each basin individually or together? I am puzzling how to calibrate soil layer depth. Based on my experience, the Noah four soil layers are 0-10 cm, 10-40cm, 40-100 cm, and 100-200 cm. The mid-layer is 5 cm, 25 cm, 70 cm, and 150 cm. If you calibrate soil depth, for each grid point at a given basin, you adjust these soil layer depths. If so, can you make a plot to compare these calibrated soil depths with default soil layer depths.

An initial sensitivity analysis was performed using one parameter at a time to identify the parameters that are sensitive to streamflow. After this analysis, we selected the parameters for calibration. Based on the prior studies (Hogue et al. 2005), we selected model parameters for the calibration. The selected parameters are tweaked manually (generating sets of model parameters and selecting the best among them based on model skill to represent observed stream flow). This is done individually for every selected river basin (see Supplement Fig. S1). We adjusted the soil depth following the (calibration) approach of the VIC model in which the soil depths are estimated via a calibration procedure such that the modeled stream-flow matches the observed values. We followed this approach given the wide success of the VIC model application in a wide variety of river basins across different climatic conditions. Moreover using a similar (calibration) approach, we aim to harmonize the different model applications over India. So, the soil depths in LSMs are treated as calibration parameters, which vary across the models and the river basins. We will make this issue clear in the revised manuscript.

As suggested, we will provide skill scores of every model using the calibrated parameters and default soil depths, which could further help illustrating the advantages of the model calibration.

(d) It is very confused how to calibrate CLM using soil depth layers. More explanations are needed.

Please refer to the above response. However, in the revised manuscript, we will discuss the calibration process in detail and the selection of the parameters.

(e) What are soil parameters in Section 2.2.2 and 2.2.3? Are they soil textures (types)? Noah and CLM use the soil textures derived from FAO, and VIC uses soil texture derived Harmonized World Soil Moisture Datbase (HWSD). I am wondering how big differences exist between two datasets? It is very well known that different texture has different soil related parameters such as field capacity, wilting point, etc., which leads to different temporal variation.

We appreciate the insightful comment from the reviewer. In the revised manuscript, we evaluated the difference in soil moisture simulations from the

VIC model using the soil parameters from the FAO and HWSD. We do not find any significant difference in these two set of parameters. The underlying (soil-textural) dataset for the HWSD is mostly derived based on the FAO datasets. Notably the different models use different pedo-transfer functions to derive soil related parameters and in this case even if the two models uses a same underlying dataset the resulting soil parameters can be different.

In the revised manuscript, we will show plots depicting similarity/differences among (common) soil parameters across LSMs.

(f) Different vegetation type classification datasets are used for different models, which can result in additional uncertainty for soil moisture product as different vegetation type has different root zone (leads to different transpiration even though surface meteorological forcing is the same).

Thanks. As stated above due to the different requirements of different models in terms of different soil and vegetation parameters, we are enforced to use different vegetation type classification datasets.

We will note this issue of additional sources of uncertainty due to requirement of different soil and vegetation datasets in the revised manuscript. However, we have evaluated the sensitivity of vegetation parameters derived from the different sources on drought assessment and results suggest no major difference.

(g) There is only one test in this study – calibrated run. I would like to see the control run/default run (the default parameters are used) and the comparison with the calibrated run. This will demonstrate what benefits you gain from the calibration process.

This is a good suggestion. For the drought assessment, the model calibration may not contribute significantly as we use only anomalies of soil moisture, which can largely be driven by the climate variations. As mentioned above, we will provide skill scores of every model using the calibrated and default parameters in the revised manuscript.

2. Model evaluation

(a) Calibrated model is only evaluated against observed streamflow. Unfortunately, I am very disappointed that the soil moisture used in this study is not evaluated against either in-situ observations or remotely sensed soil moisture. There are a few stations in India to measure soil moisture from different datasets such as Global Soil Moisture Data Bank (Robock et al. 2000), In-situ observations of soil moisture from India Meteorological Department (Unnikrishnan et al. 2016), and international soil moisture network (https://ismn.geo.tuwien.ac.at/). In addition, quite a few of remotely sensed soil moisture products such as SMAP, SMOS, SMOPS, ASCAT, AMSR2 and more are not used to evaluate LSMs soil moisture simulation products. However, the major variable used in this study is 60 cm soil moisture. Robock, A., et al., 2000: The Global Soil Moisture Data Bank, BAMS, 81, 1281-1299. Unnikrishnan, C. K., et al., 2016: Validation of two gridded soil moisture products over India with in-situ observations, J. Earth System Science, 125, 935-944.

In the revised manuscript, we will show the model skill in representing observed dynamics of soil moisture (using both the observation dataset from Global Soil Moisture Data Bank and a proxy remote sensing derived gridded soil products). We however like to note here that most of the gridded (remotely sensed) soil moisture products have their inherent uncertainty – a quite prominent one is that they are limited in their inference of soil water to only few cm from ground surface).

(b) The authors assumed 60 cm soil layer as root zone. However, for each individual model, it defines its root zone varying from vegetation type to vegetation type. For example in Noah, grass root zone is 1m and forest is 2m. I suggest the authors use 60 cm soil moisture in whole text to avoid confusing the readers.

Thank you. We will revise the text following your suggestion.

3. Model result analysis

(a) The uncertainty analysis is very limited due to three models as the samples are too few for a representative of model uncertainties. In general, the spread can roughly show an uncertainty range when three-model ensemble is used. The authors need to indicate this weakness in a discussion section.

We will include this as a limitation in the revised manuscript.

(b) The authors indicated that the uncertainty in soil moisture is mainly due to model parameterizations – resulting in different persistence of soil moisture. They assumed that there is a large field capacity for CLM but there is no further investigation. In practical, different soil texture datasets, different vegetation type classification datasets, different model structure (specific soil layer in CLM and Noah vs hydrological soil layer concept), and other ET parameterizations may affect this uncertainty together. I recommend make several sensitivity tests to clarify these issues. At least, plot field capacity, wilting point, soil type, vegetation type, root zone depth for all models and then compare their differences.

We will include plots showing differences/similarity among (common) parameters across LSMs.

(c) In Figure 3c, the seasonal cycles in Noah and VIC are comparable although the magnitude is quite different. However, that in CLM is completely different with Noah and VIC. This further suggests that soil moisture evaluation against in situ observations and remotely sensed product is needed to identify which is closer to the observations.

As mentioned above, we will include an evaluation plot for different LSMs in the revised manuscript.

(d) In line 33, page 7, the authors cited Wang et al. (2009) to explain higher persistence in soil moisture due to larger water holding capacity and thicker soil column. However, the authors used 60 cm soil layer for all models and also need plot water holding capacity for top 60 cm to verify this point.

We will include the plot of water holding capacity across different LSMs in the revised manuscript.

(e) The authors find an interesting point, that is, there are larger uncertainties in Rabi season than monsoon season. Unfortunately, the authors do not make further investigation to look for the reason. They use a general sentence "which can be associated with the role of air temperature and precipitation on soil moisture" to explain. When Figure 4 and Figure S3 are checked, during the monsoon season, three models have larger similarity than Rabi season mainly due to VIC model. A possible reason is that VIC water mode rather than energy mode is used in this study. During the monsoon season, water is unlimited and limited energy is used due to less net radiation (rainy and cloud sky). Energy and water-mode type model does not have big difference. However, during Rabi season, water is limited but energy may be unlimited, so that energy-type model (Noah, CLM) shows larger difference than water-mode type model (VIC). A quick check is to use VIC energy mode to re-run this test to compare with VIC water mode run.

Thank you for this comment. We will include this in the discussion of the revised manuscript as per your suggestion.

Minor Comments:

1. Check Table S1: Surface downward shortwave and longwave radiation , for CLM v3.0, soil texture based on IGBP or FAO or vegetation type data based on IGBP.

We will edit this in the revised manuscript.

2. Check Table S2: East coast, calibration and validation period is overlapped.

We will check this in the revised manuscript.

3. Check Table S2: Mahanadi, calibration and validation period is overlapped.

We will check this in the revised manuscript.

4. Check Table S2: Subarmarekha, calibration and validation period is overlapped.

We will check this in the revised manuscript.

I assumed that the authors used independent period to validate the calibrated models. If not, please explain the reason.

Yes we have used the independent period for the model calibration and validation. We will make this point clear in the revised manuscript.