Comments/Text of Anonymous Referee 1 (AR1) posted in blue and our text in black.

([1-11] present a thoughtful summary of our manuscript)

[12] I believe that this paper represents a very significant contribution to the Earth System literature related to the development of Dynamical Environmental Systems Models (DESMs). I have alluded to some of the problems associated with the conventional CM approach in paragraphs [2-6] above. In this regard, there has been increasing community interest in the use of both "large sample" data sets and the use of "model-structural-correction-via-data-assimilation" (learning from data) to extract better understanding about the structure and functioning of hydrological systems, such as catchments.

[13] This paper bridges the challenges of learning from large sample data sets and learning how catchments structures/behaviors can differ at local to regional scales in a very meaningful way. While not addressing the problem of prediction in ungaged basins directly, the ability of the EA-LSTM to learn from and characterize differences in catchment functioning encoded in catchment attribute data is highly significant, and it would seem that a natural next step would be for the authors to demonstrate that potential by running experiments that seek to demonstrate that predictive ability learned from gaged locations can be transferred to ungaged locations. I look forward to reading more about this in the future.

We do have a short paper on this topic in WRR, that this reviewer is also currently reviewing. That paper, however, does not explore how the catchment-aware embedding presented here as an adaptation of the LSTM architecture helps in the PUB setting. This is for future work.

[14] As such, I have only a few suggestions to offer the authors. The first is that the current title "Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling" presents a rather technical front to what is arguably (in my opinion) a much more significant piece of work. I therefore offer up the possibility for the authors to consider that the introduction and discussion/conclusions sections be somewhat revamped/broadened to reflect the perspectives offered in my above summary of the paper. As indicated, I do think this paper is really more about the interesting challenges of learning and characterizing (via dynamical systems models) the "behavior and functioning" of hydrological systems at the catchment scale in such a manner that both universal (fundamentally hydrological) principles, and local-to-regional scale uniquenesses of such systems can be learned by accessing the patterns of information encoded in large sample data sets (Gupta et al 2014). In this regard the title could also then be generalized to reflect the nature of the conversation about "Learning Universal, Regional and Local Hydrological Behaviors via Machine-Learning applied to Large Sample data Sets". Or this more general discussion could be saved for a future publication.

Thank you for the suggestion about broadening the scope implied by the title. We will take the suggestion to change the title and update the discussion accordingly. However we don't feel

confident enough to state that we already have a “universal” model, but rather that this work is a step in that direction. Thus, we would change the title to “*Towards Learning Universal, Regional, and Local Hydrological Behaviors via Machine-Learning applied to large-scale Hydrology*”.

[15] The second is that while the basin-average NSE* loss function does seem to serve the immediate needs of this study, I think that the ML-approach (and more generally hydrological learning from catchment data sets) can benefit from a more thoughtful approach to the problem of model performance metrics. In particular, the use of the observed output data “mean” as a benchmark for constructing the NSE itself, and the use of the output data variance to “normalize” across catchments to obtain somewhat comparable metric values to be averaged (or otherwise summarized in some statistical manner) seems, to me, problematic. In this regard, I think an Information Theoretic approach might ultimately prove to be more meaningful. I point out that the value of the metric, when used as the basis for assessing across different catchment locations, would be much enhanced if it somehow recognized the relative differences in complexity/difficulty associated with modeling the dynamical input-state-output behaviors at different locations (due to climatic, geological, and other factors). As discussed by Schaefli and Gupta (2007), the problem is at least partly one of appropriate benchmarking in order to make metric values meaningfully comparable. Some types of catchments (such as humid ones perhaps) are relatively easy to model to the level of obtaining high performance (e.g. NSE) values, while others (such as arid ones perhaps) are much more difficult to model … potentially requiring more complex model structures, more data, and perhaps better data quality. Since the challenge here is learning hydrological principles from the data, and some catchment systems are easier to characterize using simpler model structures, it would seem prudent to figure out how to account for this knowledge in the designs of our learning systems, which includes the metrics used as the filter through which information is being extracted.

We absolutely agree that it will be critical, going forward, to understand carefully and in detail how different loss functions affect the training of deep learning Hydrology models. We’ve done some work with this - trying to emphasize peak and low flows, working with probabilistic loss functions, etc. None of that work is mature enough to publish at this point. This has been a much bigger challenge that we perhaps originally expected, and we appreciate the reviewer’s advice. Hopefully we will have something more meaningful or helpful to say about this in a future publication.

[16] Finally, I think that the aforementioned issue may also relate to the fact that certain catchment attributes tend to be dominant indicators of differences in catchment behaviors, while others seem to show “lower importance” (sensitivity). It is been well known that “climate” (and one would reasonably expect also “topography”) is the dominant indicator of catchment similarity, but this does not really help us to understand what structural differences in catchments drive differences in their behaviors. The finding that soil and vegetation characteristics are low on the “importance” list is interesting, as it suggests that the existing catchment attributes being used may not be sufficiently informative about catchment-scale soil and vegetation contributions to hydrological behaviors. So, is it a problem of poorly encoded

soils and vegetation information at the catchment scale, or is really the case that such soils and vegetation do not play as big a role in hydrological behavior as we might expect? It would be interesting to consider how this issue could be better investigated using the ML approach.

First we would like to clarify that we do not say that soil and vegetation indices are not "important" but just that climatic and topographic attributes are *more* important. As the reviewer mentions, this follows hydrological literature and also the intuition of the reviewer. In our case, this could potentially result in two basins with similar climatic and topographic attributes that are distinguished primarily by their soil/vegetation properties. However, in the larger context it is the set of climate and topographic attributes that "separates" the most basins and thus the larger sensitivity/importance.

Regarding the question about what we might be able to learn from these results for hydrological modeling. This seems in-line with the well-known idea that the first-order trends in most models are Budyko-type effects (i.e., climate related). This is not especially new, but also encouraging that the LSTM behaves as expected. One thing we might take away from this is due to the comparison with MPR regionalization. MPR uses topographic, geologic, soil and land use attributes as inputs. We show that our regionalization approach outperforms the two MPR calibrated models (VIC and mHM). This could indicate that (the conventional use of) MPR might be improved by including climatic attributes in the regionalization scheme.. I guess the question is about the extent to which it is meaningful for a model to 'react' directly to climate indexes rather than just to meteorological forcings. The regressions in typical regionalization strategies could use climate indexes as regressors or inputs.

Another point regarding the relative unimportance of geological and vegetation features indicated in our sensitivity analysis is that probably catchment averaged soil properties, as well as vegetation indices contain too much noise compared to the relatively noise-free climatic and topographic attributes. This is what is probably meant by the reviewer with "poorly encoded" information of those features, in which case we would agree. We don't, however, prove this in the paper - all we show in the paper is that there is enough information in the indexes to at least help with catchment differentiation, and that traditional approaches do not utilize all of this information.