Reviewer #1, Dr. Christa Peters-Lidard
General Comments:
The general objective of this work is to extend and generalize the diagnostic soil moisture model of Pan et al, 2003 and Pan 2012 by recasting the work into an hourly time step and calibrating the parameters of the model using SCAN data. The authors then explore how hydro-climatic and edaphic similarity can be used to transfer parameters between sites.

This work is generally well described and relevant to HESS readers. My biggest concerns are the presentation of results, and particularly the lack of consistency in the evaluation metrics among the various analyses. The manuscript really needs a table showing site number, site characteristics such as hydroclimate class and soils, as well as different error metrics (ubRMSE, Bias and R^2) before and after bias correction.

We appreciate the reviewer's timely reply and positive impression of our work. We have worked to improve the charts and tables, along with the descriptions thereof, in their captions and surrounding text. The table described in the paragraph above, one which encompasses the fifteen sites used, their hydroclimatic class, their general soil characteristics, and their performance before & after bias correction has been constructed, now Table 1. This table is then referenced on lines 357-360 (after p. 2335, line 3 in the HESS-D paper).

> "Table 1 presents all fifteen sites for which the diagnostic soil moisture equation has been calibrated, including information regarding their hydroclimatic class from Coopersmith et al (2012), their soil textural characteristics, and their performance before and after the KNN bias correction process."

SiteID	Hydro-climate	Soil Information	RMSE	RMSE w/ KNN	R^2	R^2 w/ KNN
2008	LJ	Sandy Loam	8.38	7.69	0.590	0.726
2013	LWC	Sandy Loam	2.16	2.06	0.876	0.885
2015	IAQ	Loamy Sand	3.29	2.37	0.740	0.841
2017	ISQJ	Sandy Loam	3.62	3.27	0.637	0.701
2018	IAQ	Loamy Sand*	2.23	2.16	0.803	0.828
2028	LPC	Loam	4.89	4.71	0.707	0.738
2031	ISQJ	Silty Clay Loam	5.46	6.00	0.687	0.750
2036	LPC	Silt Loam	4.61	3.95	0.635	0.726
2038	LJ	Sandy Loam	4.81	4.51	0.546	0.584
2068	ISCJ	Silty Clay Loam	5.28	4.03	0.716	0.837
2089	LJ	Sandy Loam	6.7	6.31	0.682	0.697
2091	LPC	Silt	8.12	6.89	0.539	0.808
2107	IAQ	Loamy Sand	1.98	1.85	0.790	0.843
2108	IAQ	Loamy Sand/Sand	1.26	1.12	0.828	0.863
2111	ISQJ	Silty Clay Loam	5.38	5.01	0.607	0.796

*Not similar to other sandy soils, see Figure 12.

Table 1, The Fifteen SCAN Sites: Class & Soil Information and Performance

The results from the k-NN bias correction procedure shown in Figure 2 show only the total R^2 before and after correction, and only by site number, rather than by something related to the hydro-climatic or edaphic similarity (see suggestion for table above). Further, the importance of the bias correction is likely a seasonal or diurnal bias, and this should also be discussed/shown, perhaps as separate lines on Figures 3-6 or as an add-on to Figure 2 for different types/ magnitudes of bias correction. Are the sites shown in Figures 3-6 representative?

The purpose of the initial k-NN bias correction is to demonstrate the utility of that technique in improving the results of the diagnostic soil moisture equation rather than the importance of hydroclimatic or edaphic similarity. Essentially, these figures are intended to demonstrate the utility of our enhancements to the diagnostic soil moisture equation by applying those techniques to results calibrated at these sites. Regarding the importance of the bias correction in seasonal or diurnal terms, this discussion is broached with Figure 6, but should be expounded upon. In Figure 6, the improved 'shape' of the ML-enhanced soil moisture series relative to the observed measurements in blue explains, in part, the improvement in correlation observed. However, as the green line is also nearer to the actual observations than the original diagnostic soil moisture equation results (red), the improvement extends beyond the ability to introduce a diurnal cycle. The new Figures 7, 8, and 9 present the average bias correction as a function of the hour of day, the day of year, the soil moisture value estimated, and the beta series from the diagnostic soil moisture equation (the summed, decaying precipitation series). This allows a deeper assessment of how this machine learning approach is improving soil moisture estimates, forming a new section (lines 389-418) of the revised manuscript.

> "Figure 7, 8, and 9 present these results in more detail for each of the three SCAN sites presented in Figures 3, 4, and 5. In each figure, the upper-left image presents the average bias correction (change in % soil moisture) for each hour of the day (0-23). At all three sites, bias corrections display a clear diurnal pattern – that is to say the removal of a diurnal cycle is a substantial role of machine learning under a variety of hydroclimatic and edaphic conditions. The upper-right image of each figure presents the bias correction as a function of the unadjusted soil moisture estimate – essentially, whether there exists a systemic over- or underestimation when values are high or low.
>
> The first two sites (Figures 7 and 8) do not present a clear pattern, but Figure 9 displays a trend suggesting that the highest estimates of soil moisture tend to be overestimates and the lowest estimates of soil moisture tend to be underestimates – but these biases are removed via machine learning. The lower-left image presents bias correct as a function of the day of the year (from 100-300, the days of the year when the model is applied). At all three sites, the seasonal cycle does appear in terms of the patterns of bias correction, but the pattern is noisier than the diurnal cycle. The magnitude of the adjustments are largest in the monsoon-affected desert of New Mexico, a bit smaller in the Midwestern plains characterized by less extreme seasonal behavior, and smallest in the Southeast where seasonal variations are low.
>
> Finally, the lower-right image relates bias correction to the beta series from the diagnostic soil moisture equation (Pan, 2012), a convolution of a decaying precipitation time series working backwards temporally from the current time. Stated differently, these charts relate bias correction to the amount of antecedent precipitation (with more recent precipitation weighted more heavily). In Figure 7 (Plains, Silty Clay Loam), the model tends to underestimate moisture when large quantities of antecedent rainfall are present, where in Figure 9 (Woods, Sandy Loam), once antecedent precipitation becomes non-trivial, displays the opposite pattern. This is consistent with the finer Midwestern soils' proclivity for ponding/flooding due to larger proportions of clay. In these cases, larger amounts of rain will soak soils from above, and capillary rise might further soak sensors from below, leading to underestimation from the diagnostic soil moisture

equation and subsequent machine learning correction. By contrast, with sandier soils, drainage occurs easily, leading to higher rates of loss than the eta series (Pan, 2012) would predict (there is more available water to lose), leading to overestimation with large amounts of antecedent rainfall.”

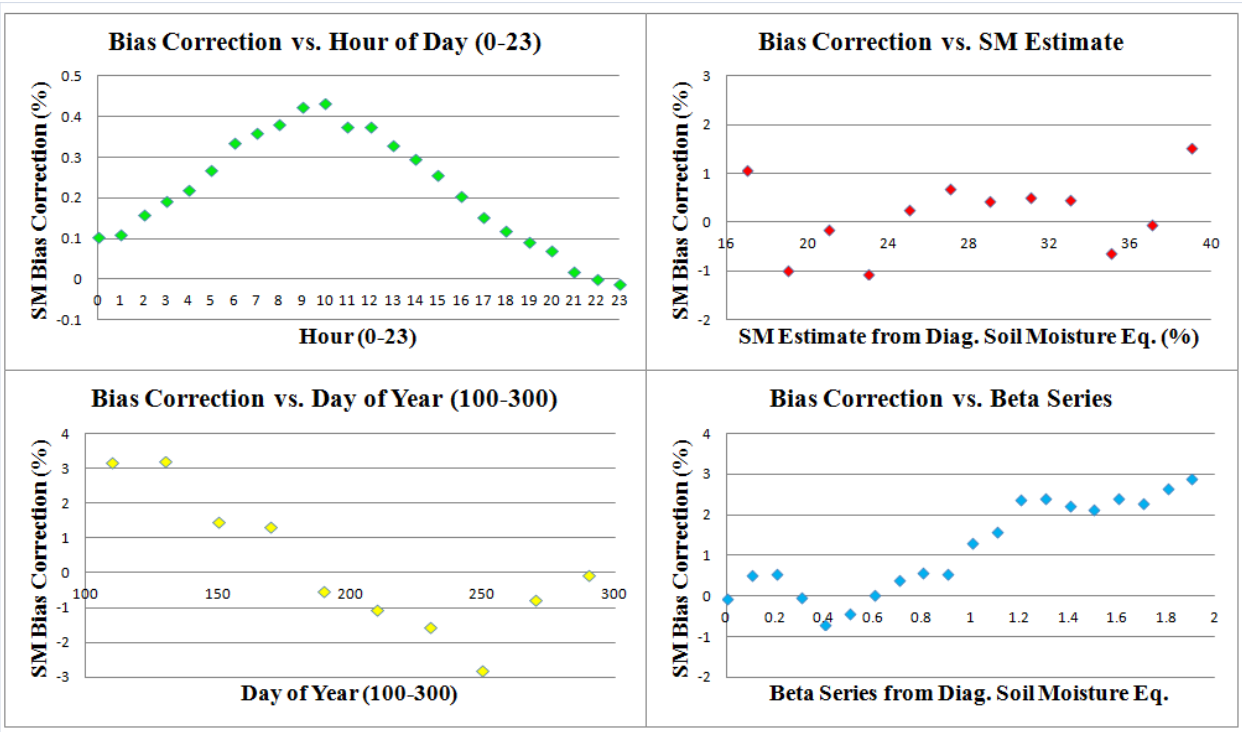

Figure 7, Bias Correction Analysis, SCAN Site 2015 (IAQ, Desert, Loamy Sand)

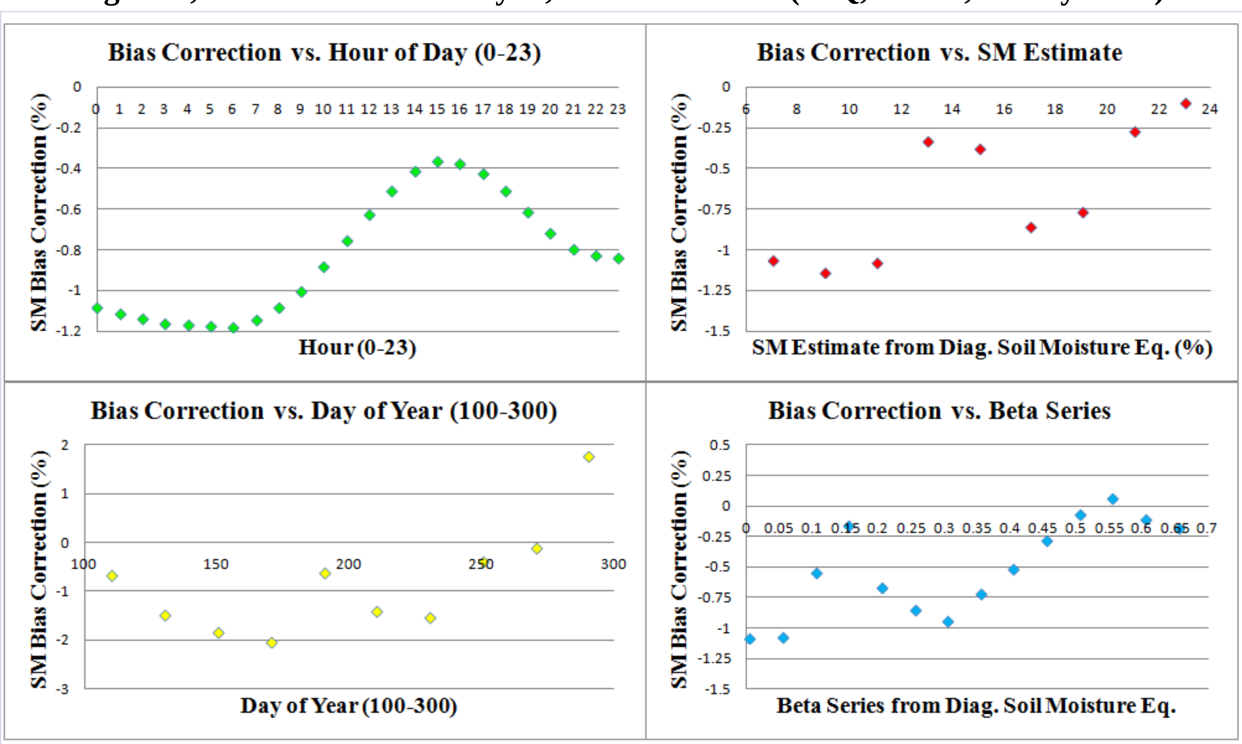

Figure 8, Bias Correction Analysis, SCAN Site 2068 (ISCJ, Plains, Silty Clay Loam)

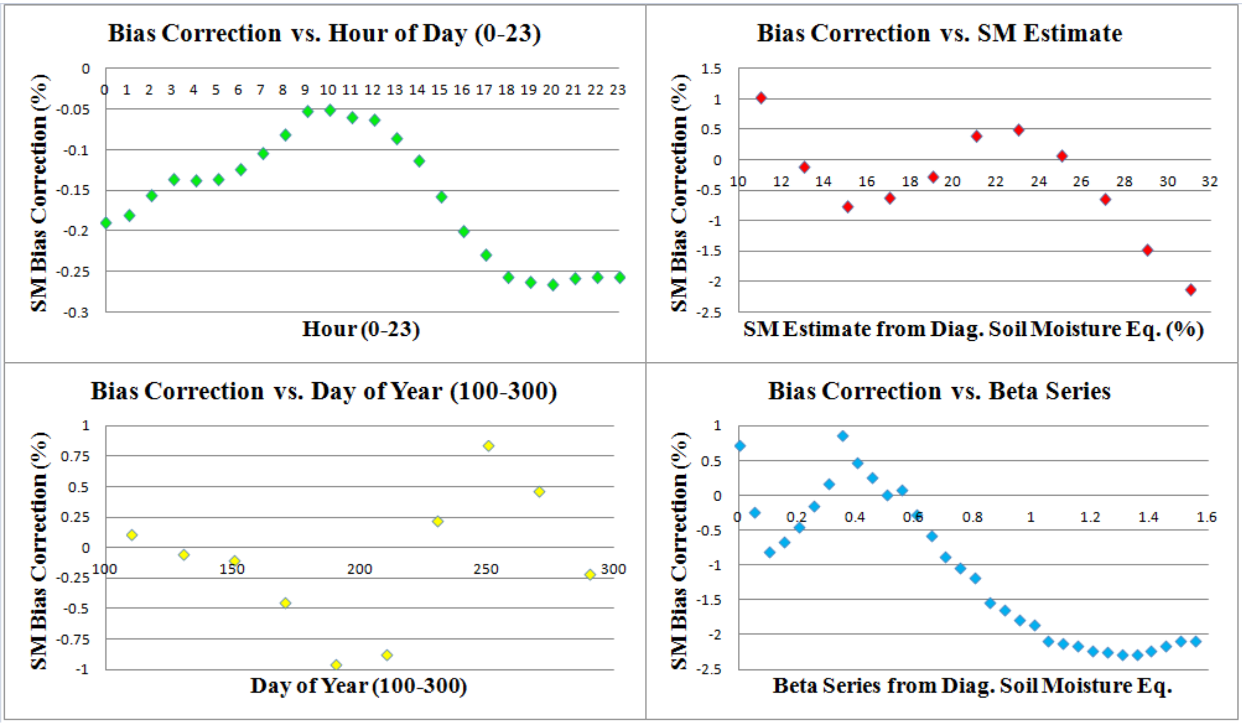

Figure 9, Bias Correction Analysis, SCAN Site 2013 (LWC, Woods, Sandy Loam

In terms of the representation of the three catchments presented in Figures 3-5, they fall within three different hydroclimatic classes (IAQ, ISCJ, LWC), within three different regions of the continent (Southwest, Midwest, East), with differing soil types (Loamy Sand, Silty Clay Loam, Sandy Loam), and display improvements with respect to machine learning that are roughly consistent with the improvements shown over the fifteen sites examined (expanded upon on lines 323-334 of the edited manuscript). Figure 6, a zoomed-in version of Figure 3, is admittedly a more extreme example of a diurnal cycle (the conditions are very warm and dry in New Mexico, and the fluctuations are large relative to the average soil moisture value) – it was chosen primarily to illustrate the effect discussed rather than to imply that such cycles are always the dominant source of variance relative to the traditional wetting/drying processes.

> "To explore these findings in more detail, three of the 15 SCAN sites, chosen to represent different hydro-climatic locations – New Mexico (#2015, hydroclimate IAQ/southwestern desert, Loamy Sand), Iowa (#2068, hydroclimate ISCJ/northern midwest plains, Silty Clay Loam), and Georgia (#2013, hydroclimate LWC/southeastern forest, Sandy Loam) are examined to illustrate how improvements from adding machine learning error models to the diagnostic soil moisture equation differ across sites. These three sites represent three distinct hydro-climatic classes, with significant differences in soil texture, seasonality of precipitation, aridity, timing of maximum precipitation, and timing of maximum runoff. Using error correction models for prediction at these sites increased R^2-values by an average of 8.2%, which is similar to the 8.3% improvement in R^2 averaged across all fifteen sites. Thus, these three locations are representative in terms of both hydro-climatic and edaphic diversity and their responsiveness to machine learning."

The most important results in the work are the results shown in Figure 7/Table 1 as well as the Venn diagram in Figure 11. However, similar to the suggestion above, it would be useful to examine these cross-validation results with a common set of statistics as with the initial calibration results. The boxplots do give a sense for the distribution of errors, but the description of what is actually shown in both Figure 7 and Table 1 is a bit confusing. The captions need more information across the board. Is the loss of Rˆ2 equivalent to the simple difference in Rˆ2 (baseline-new)?

Yes, the difference in R^2 between the baseline values and new (bias-corrected) values is the same calculation shown in Figure 7 and Table 1. We apologize for the confusion and concur that applying a consistent metric is important – we have done so, but need to ensure that this is clear to the reader. Additional explanation appears on lines 407-409 of the edited manuscript to clarify what has been done.

> "Figure 10 presents box plots illustrating the change in R^2 values for these three sets of pairs in a manner analogous to the differences shown in Figure 2. Table 2 presents the quantitative results, again averaging the deterioration of performance in terms of change in R^2."

The results for "similar" sites are interesting but what exactly constitutes a "similar" site is tersely defined with the discussion of Figure 7: "(different by a single split within the classification tree)." At a minimum you should refer back to Figure 1, where presumably the components of the tree algorithm are derived using MOPEX data.

With the new table constructed (Table 2, discussed in the response to the reivewer's first point), it should be clearer which sites are considered to be within the same hydroclimatic class (column #2) or within the same edaphic group (column #3). In terms of figure 7, the yellow box-plot, referring to 'similar' hydroclimatic classes (different by one-split on the classification tree), additional explanation is certainly required. In this case, ISCJ/ISQJ, LWC/LPC, ISQJ/IAQ, ISCJ/IAQ, LJ/LPC, and LJ/LWC are considered 'similar,' as in all of these pairs, tweaking a single hydroclimatic indicator could move the catchment from one class to the other. An additional note is added to indicate that the classes chosen are those determined by more recent data, as the original classes from Coopersmith et al (2012) have undergone climate change since their original classifications using MOPEX data from 1948-2003 (Coopersmith et al, 2014).

Overall, I think the work is worthy of publication, and just needs some moderate revision to standardize the statistics and also expand captions and discussion to make explicit links and provide better explanations of the results being shown.

We appreciate the reviewer's endorsement and hope the changes made will be satisfactory.

Specific Comments:
Figures 3-5 do not have legends, and appear to be cut off on the right side. The legend on Figure 6 is barely legible. The captions don't need to be repeated. You could say "same as Figure 3 but for site xxxx, which is a <fill in type> hydroclimate and <fill intype> soil."

Regarding the right edge, this is an idiosyncrasy of the margin settings within the manuscript file sent to HESS – this has been corrected. Legends are now included and/or enlarged. The captions have been altered to provide the full description in the first of the three figures and refer back to Figure 3 thereafter. Hydroclimate information and soil textural characteristics are now listed in all cases.

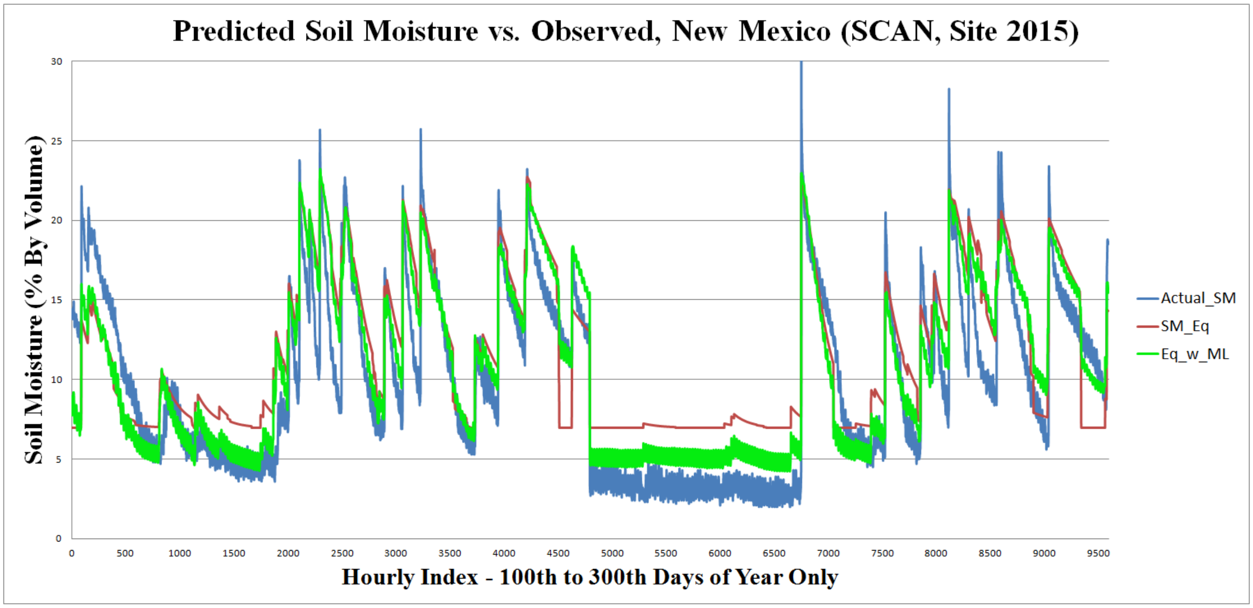

Figure 3, Soil Moisture Time Series, SCAN Site 2015, New Mexico (USA), Actual Soil Moisture (Blue Line), Diagnostic Soil Moisture Equation Estimate (Red Line), and Diagnostic Soil Moisture Equation with Machine Learning Error Correction (Green Line). Hydroclimate: IAQ (Intermediate Seasonality, Arid, Summer Peak Runoff)
Soil Texture: Loamy Sand

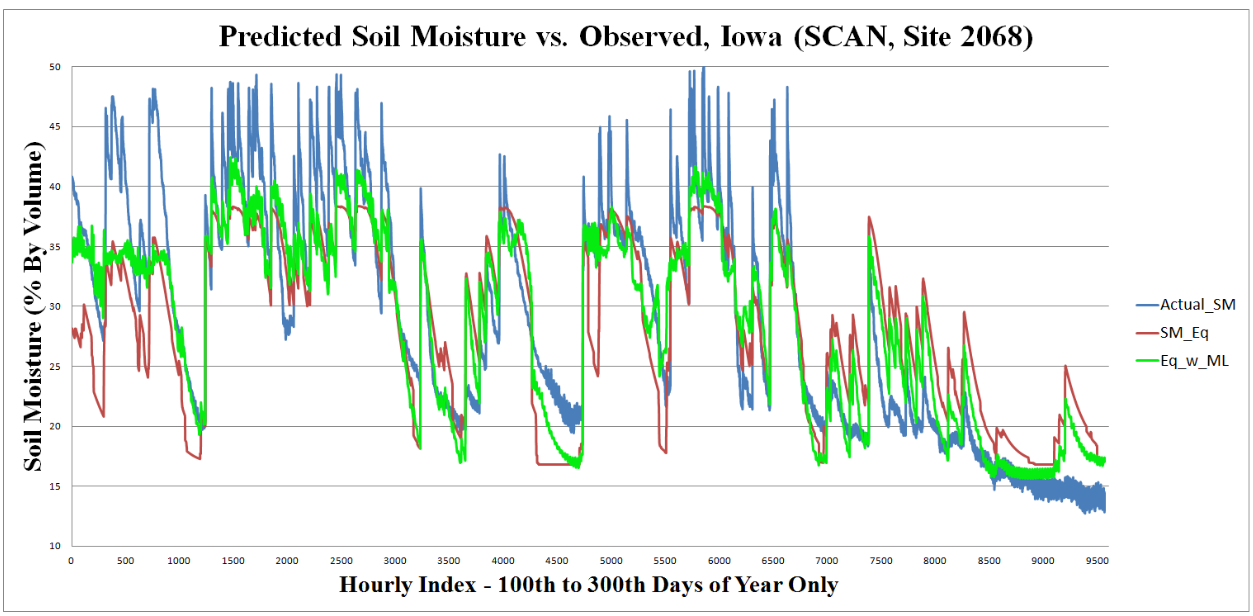

Figure 4, SM Time Series, SCAN Site 2068, Iowa (USA), line colors from Fig. 3
Hydroclimate: ISCJ (Intermediate Seasonality, Semi-Arid, Winter Peak Runoff, Summer Peak Precipitation)
Soil Texture: Silty Clay Loam

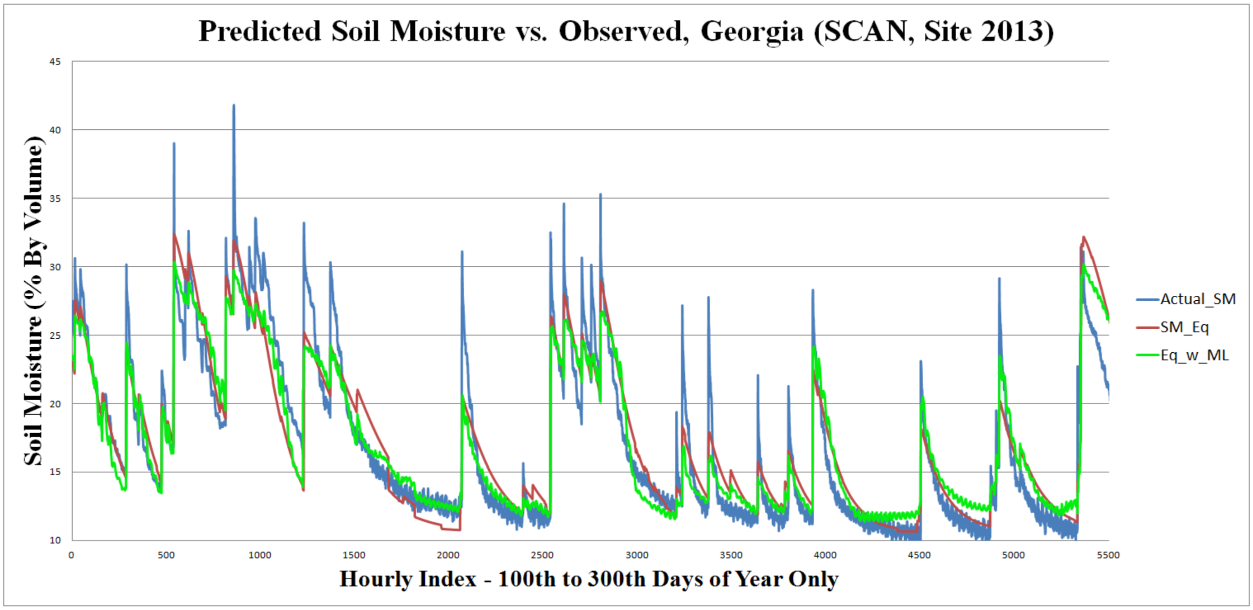

Figure 5, SM Time Series, SCAN Site 2013, Georgia (USA), line colors from Fig. 3
Hydroclimate: LWC (Low Seasonality, Winter Peak Precipitation, Winter Peak Runoff)
Soil Texture: Sandy Loam

Figure 9 and Figure 10 are redundant. Only Figure 10 is necessary.

Agreed, Figure 9 has been removed.

Figure 11 is a nice summary of the net effect, but I'm left wanting to see more results for the four cases illustrated in the diagram. I would think the type of errors due to soil and hydroclimate are different. Some more insightful discussion here is warranted.

This is an important question, but one best left for future work. Lines 515-519 of the revised manuscript discuss this question in further detail.

> "It is likely that the types of errors made when parameters are cross-applied between sites of different hydroclimates will differ from the types of errors that appear when the sites differ edaphically. Further research into the specific conditions under which models err, along with the magnitude and bias of those errors, would be highly useful."

When one begins to analyze types of errors under various cross-applications, the computational complexity reaches a level that suggests a separate inquiry would be required. Essentially, there would be 210 (x,y) pairs to evaluate in detail – this would make for an excellent follow-up paper, but thecomputational complexity of understanding how ~40,000 examples per site calibrate roughly 10,000 examples in 210 cases with extensive analysis regarding how and why the error occur would likely fall beyond the scope of this paper.

Technical Issues:
Throughout the manuscript, the references to Pan et al., 2012, need to be corrected to Pan 2012, as there are no other coauthors.

Good catch. The relevant changes have been made.