Responses to the comments from Reviewer #1

We thank the reviewer for the constructive comments. We have incorporated the review comments and revised the manuscript thoroughly. The review comments and the revision have resulted in a much more complete presentation of the work. While the changes made to the manuscript can be seen in the revised manuscript, we also present here our detailed responses to the review comments (reviewer comments in black, our response in blue).

GENERAL COMMENTS

The paper describes a very simple approach to attribute model biases in the simulated states and fluxes of the latest version of the Community Land Model (CLM4.5). This is an important and interesting research area, as biases in modelled soil moisture or discharge can for instance substantially affect the prediction and analysis of hydro-climatic extremes such as droughts and/or floods. The approach introduced in the paper is not really innovative as it was first published by Parr et al. in 2015; but it is tested here for a larger study area and a different land-surface model. In general, the method and the results in this paper are well-described, but–to my opinion–not really surprising and rather straightforward. Substantial parts of the results and discussions are dedicated to the differences in bias between the GLEAM-derived datasets and the CLM-runs with and without the bias correction. These results are very straightforward and predictable, as the bias-correction factors were first calibrated against GLEAM. Furthermore, most of the validations/comparisons are performed at aggregated variables (both in space and time), which might mask some of the potential issues. Summarized, I think the topic of this study is interesting, but I have the feeling that the paper (especially the results section) needs some improvements before final publication. Below I list some more specific comments.

SPECIFIC COMMENTS

1. In Section 4.2.1 it is claimed several times that the performance of CLMET is substantially better as compared to the original CLM. To my opinion, these statements need to be revised as they are not necessarily correct; especially not when the reference data is the GLEAM dataset itself. As the bias-correction factors are calculated using the GLEAM data as a reference, it makes perfect sense that applying these correction factors in the model brings the model closer to GLEAM (unless the assumption of time-invariance would not be fulfilled). Therefore, the results discussed from P11-L243 to P14-L305 (i.e. comparison of the bias-corrected CLM evaporation to the GLEAM dataset) only show the robustness of the correction factors. They do not show an improvement of CLMET in reference to CLM. To me, the evaluation of the runoff coefficients and the comparison against alternative datasets of evaporation (FLUXNETMTE, MODIS) is a step in the right direction, but only a small portion of the discussion is dedicated to these results. Therefore, I would suggest to improve the evaluation of the results to really show the impact of applying the method. I would strongly recommend to (1) validate the modelled evaporation against in situ

measurements (for instance data from single eddy-covariance towers) and, (2) extend the evaluation of the model against the alternative datasets of evaporation.

Response: We have followed the reviewer's suggestions in revising the manuscript:

1) Validate the modelled evaporation against in situ measurements:

We selected 16 eddy flux tower stations from the AmeriFlux network to validate model performance (as shown in Figure 1b of the revised manuscript). These stations were previously used to validate the NLDAS-2 surface models by Xia et al. (2015). The 16 stations are located in different sub regions of CONUS with different vegetation cover (i.e., grassland, cropland, needleleaf forest, broadleaf forest, and mixed forest). Considering both consistency in validation period and data availability, we use the year of 2005 for validation at most sites except for three sites: Sylvania Wilderness (2002), Donaldson (2004) and Walnut River (2004).

The model validations are based comparing each station with the model grid cell that encompasses the station. The station-based ET (or latent heat flux, in W/m2) are measured every 30 minutes and aggregated to daily and monthly values. Except for Port Peck and Wind River Crane stations in the northwest CONUS, for all other stations the monthly mean ET from CLMET agrees better with the observed ET than that from CLM (Figure 8 of the revised manuscript). The same statement holds for daily mean ET (Figures 9, and 10 of the revised manuscript). Generally, CLM overestimates ET as compared with station observations, and CLMET alleviates this overestimation, which is consistent with comparisons between modelled ET and satellite-based ET products.

"*In addition, the ET validation is also conducted on the site scale (Figures 8, 9, and 10).* Except for Port Peck and Wind River Crane stations in the northwest CONUS, for all other stations the monthly mean ET from CLMET agrees better with the observed ET than that from CLM *(Figure 8). The same statement holds for daily mean ET (Figures 9 and 10). Generally, CLM overestimates ET as compared with station observations, and CLMET alleviates this overestimation, which is consistent with comparisons between modelled ET and satellite-based ET products.*" (last paragraph of Section 4.2.1 in the revised manuscript)

Xia, Y., Hobbins, M. T., Mu, Q., & Ek, M. B. (2015). Evaluation of NLDAS-2 evapotranspiration against tower flux site observations. Hydrological Processes, 29(7), 1757-1771.

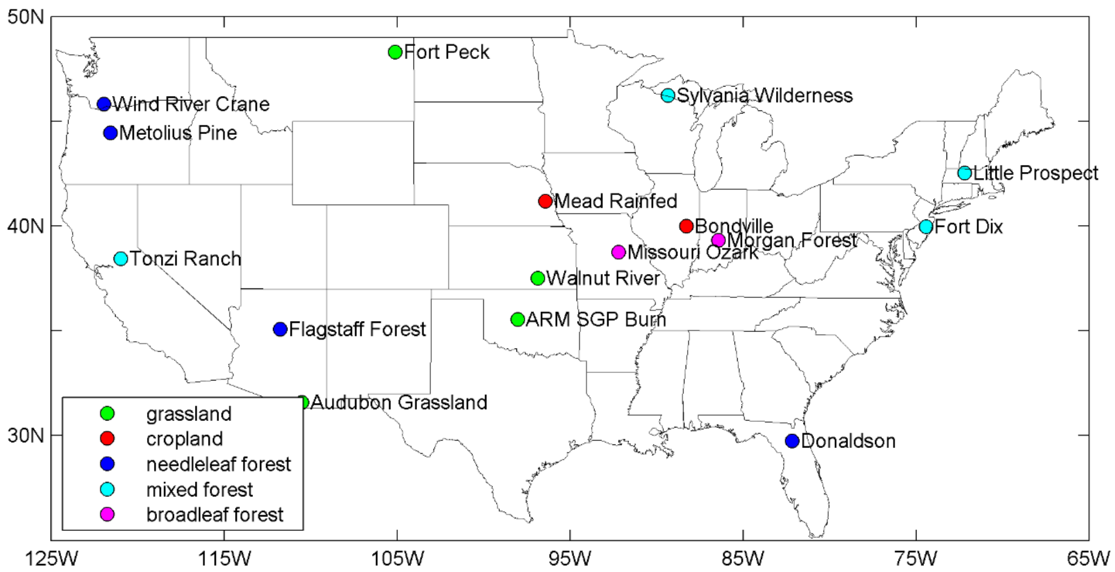

Figure 1b Locations of the 16 AmeriFlux stations with vegetation types.

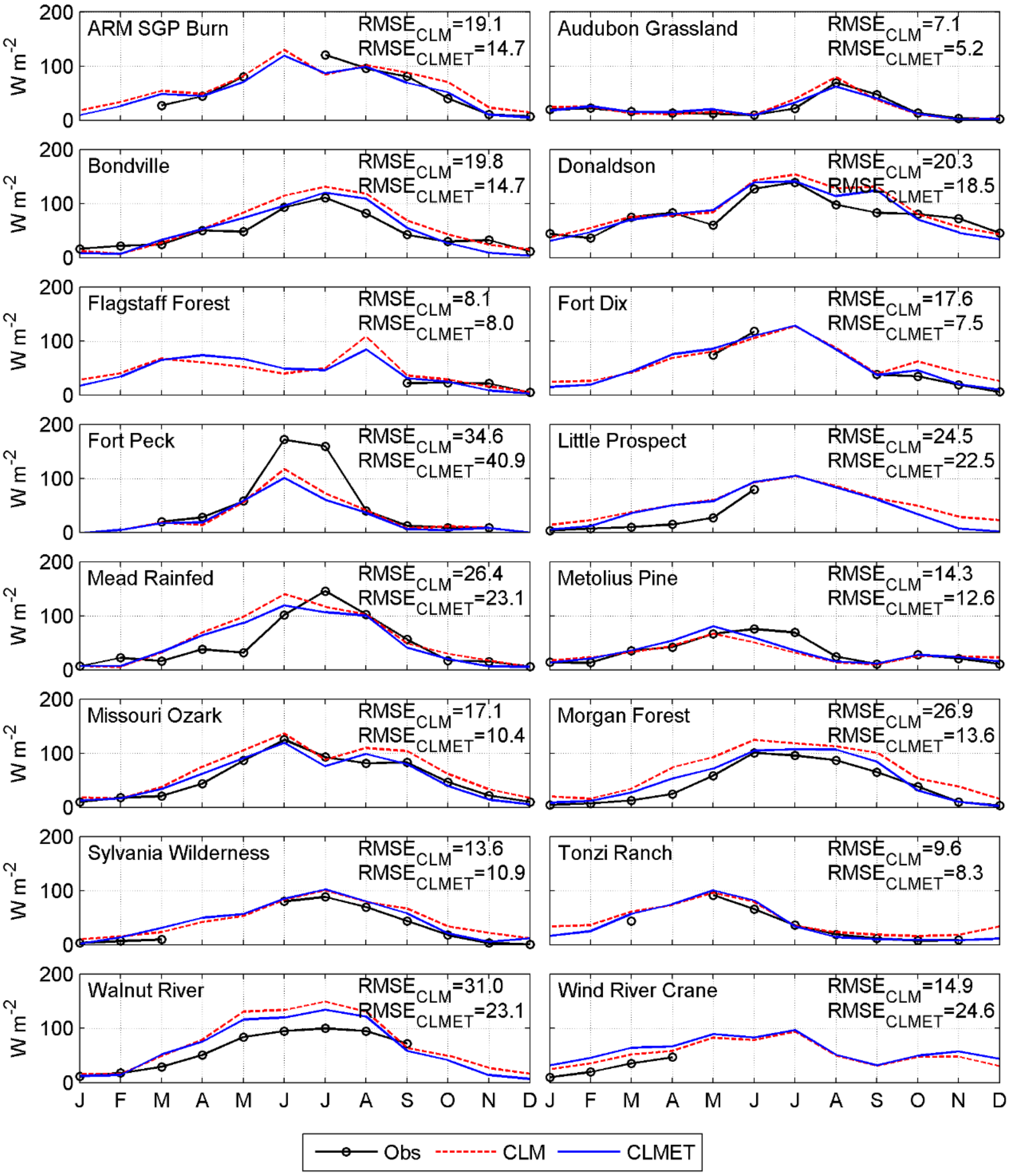

Figure 8 Monthly mean latent heat fluxes from CLM, CLMET and observations at 16 flux tower sites. RMSE$_{CLM}$ and RMSE$_{CLMET}$ represent the root mean square error against observations for CLM and CLMET, respectively. Note that the CLM and CLMET simulations are driven with meteorological forcing at the grid cell level (as opposed to site-specific forcing).

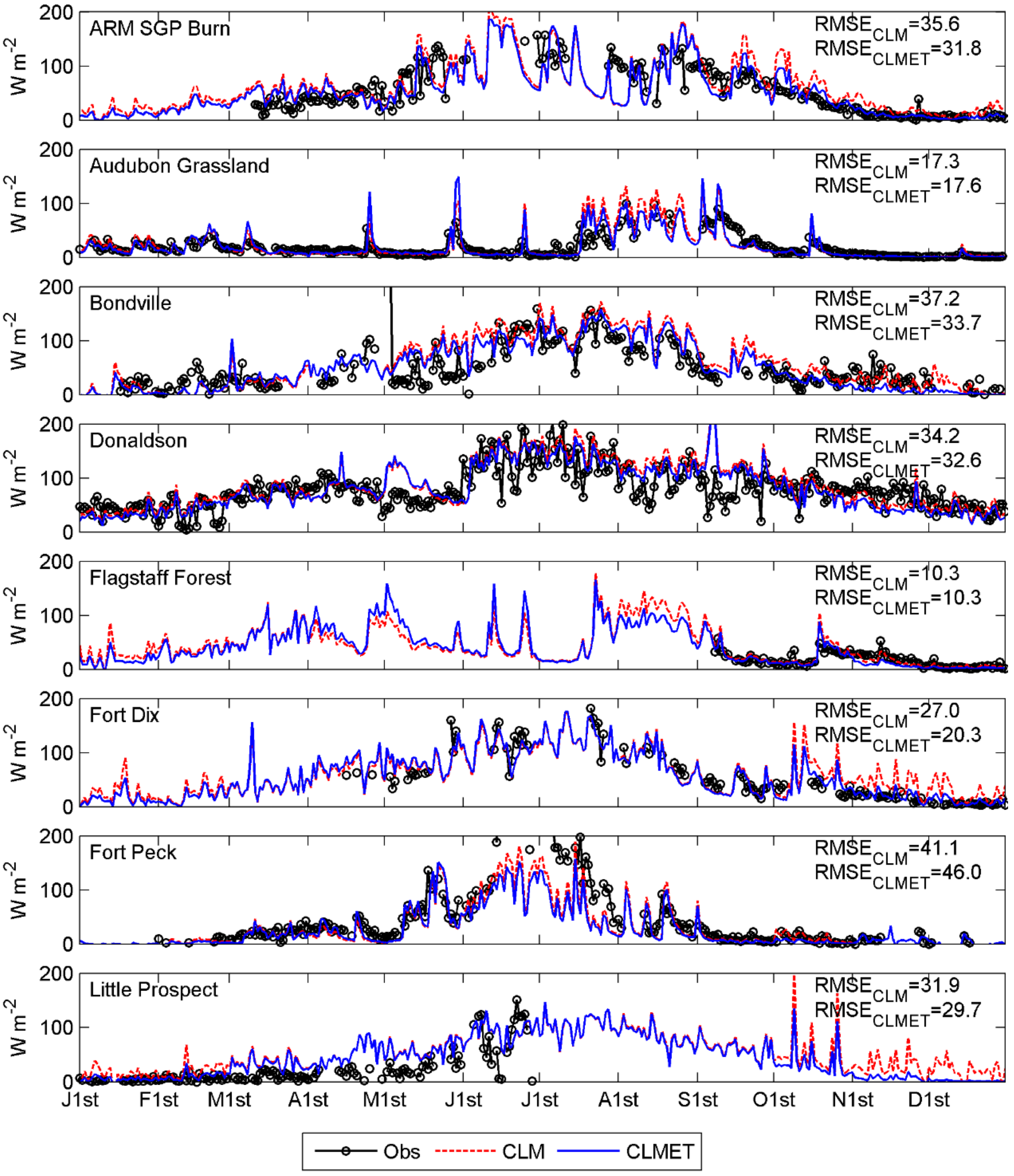

Figure 9 Daily mean latent heat fluxes from CLM and CLMET grids and station observations at ARM SGP Burn, Audubon Grassland, Bondville, Donaldson, Flagstaff Forest, Fort Dix, Fort Peck, and Little Prospect. RMSE$_{CLM}$ and RMSE$_{CLMET}$ represent the root mean square error against observations for CLM and CLMET, respectively.

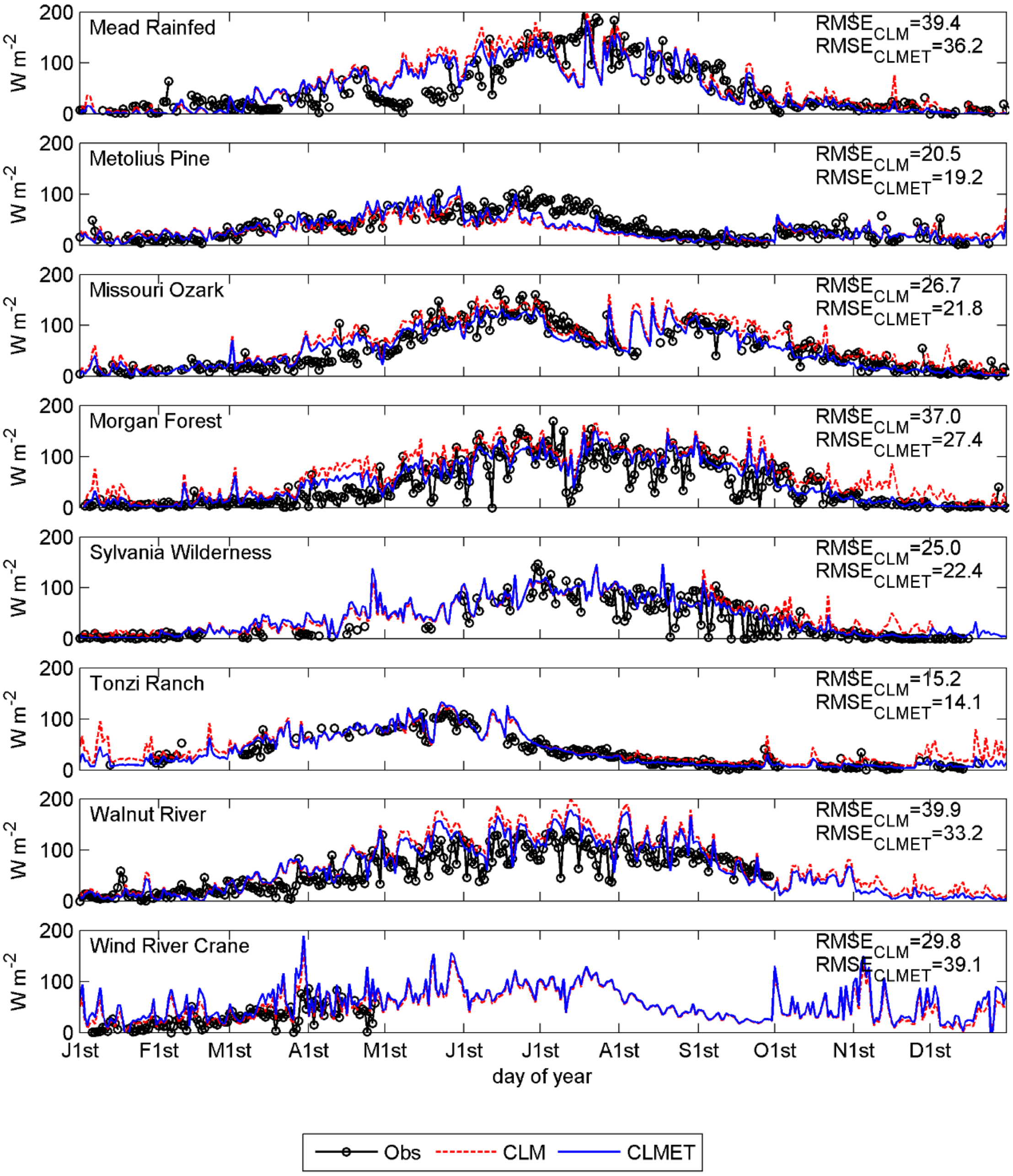

Figure 10 Daily mean latent heat fluxes from CLM and CLMET grids and station observations at Mead Rainfed, Metolius Pine, Missouri Ozark, Morgan Forest, Sylvania Wilderness, Tonzi Ranch, Walnut River, and Wind River Crane. RMSE$_{CLM}$ and RMSE$_{CLMET}$ represent the root mean square error against observations for CLM and CLMET, respectively.

2) extend the evaluation of the model against the alternative datasets of evaporation:

We have deleted the evaluations of ET seasonal cycle and monthly value using the GLEAM dataset, and added the evaluations using the MODIS and FLUXNET-MTE dataset. Therefore GLEAM is used for algorithm calibration while the other two ET products are used for validation. Using MODIS or FLUXNET-MTE ET as a reference, modeled ET from CLMET is the similar to that from CLM over western CONUS, whereas CLMET substantially improves ET simulations over eastern CONUS as compared with CLM. The improvement in CLMET is more evident during September-October-November. We have added the figures (Figures 6 and 7 in the revised manuscript) and revised the relevant texts in the revised manuscript.

"*The analysis on time series of ET from MODIS, FLUXNET-MTE, and two types of simulations also demonstrates improvement from CLM to CLMET. Climatological seasonal cycles of ET over CONUS and four sub regions for 2000-2011 are shown in Figure 6. CLMET performs better than CLM over CONUS with smaller RMSE (0.31 versus 0.40 against MODIS, 0.19 versus 0.25 against FLUXNET-MTE). The improvement mainly results from reduction of overestimation existing in CLM for SON and DJF. However, the model performance greatly varies with region. As indicated by the ET RMSE values, CLMET and CLM perform similarly over western CONUS, whereas CLMET improves the ET simulation over eastern CONUS no matter which reference data is used. Figure 7 compares the temporal evolution of the simulated ET in CLM and CLMET against MODIS and FLUXNET-MTE ET over CONUS and four sub-regions. It is evident that the bias correction method in CLMET is very effective in reducing overestimation (positive bias), but does not work as well in correcting the underestimation (negative bias). The difference has to do with the specific ET regime, i.e. whether ET is limited by water or energy. When an overestimated ET is overwritten with a lower value, the water on land is sufficient to support the reduced ET; in contrast, when an underestimate ET is overwritten with a higher value, the land surface model checks whether water storage in soil layer and vegetation canopy can sustain the elevated ET and further adjust if necessary to keep with the mass conservation equation. The extent to which ET increases is limited by the availability of water stored in soil layer and vegetation canopy. Therefore, in case of water-limited ET, the actual ET after the water availability check in CLMET can be substantially lower than the corrected ET fed into model.*" (the second paragraph from bottom of Section 4.2.1 in the revised manuscript)

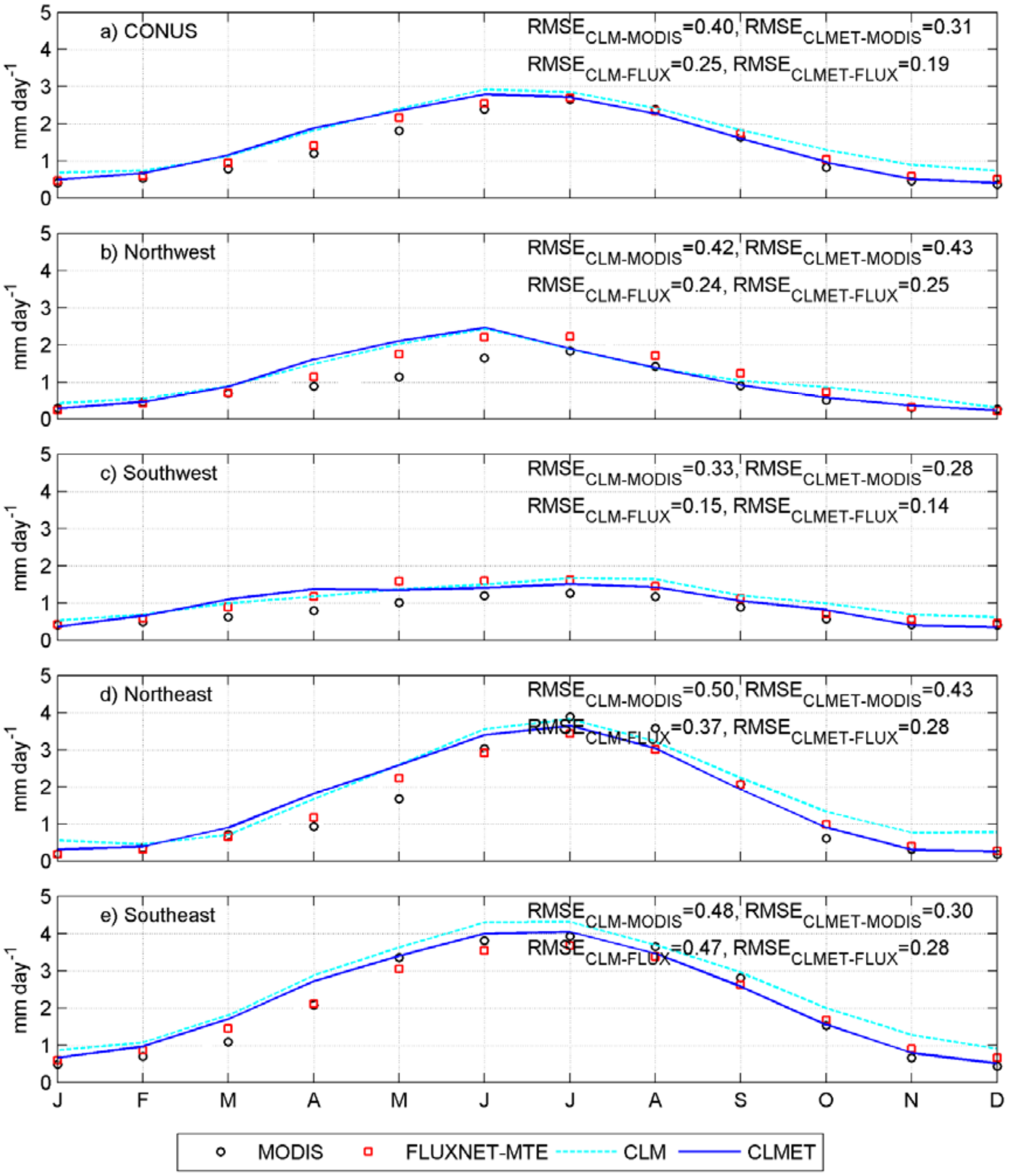

Figure 6 Seasonal cycles of ET from MODIS, FLUXNET-MTE, CLM, and CLMET over CONUS, Northwest, Southwest, Northeast, and Southeast during the period 2000-2011.

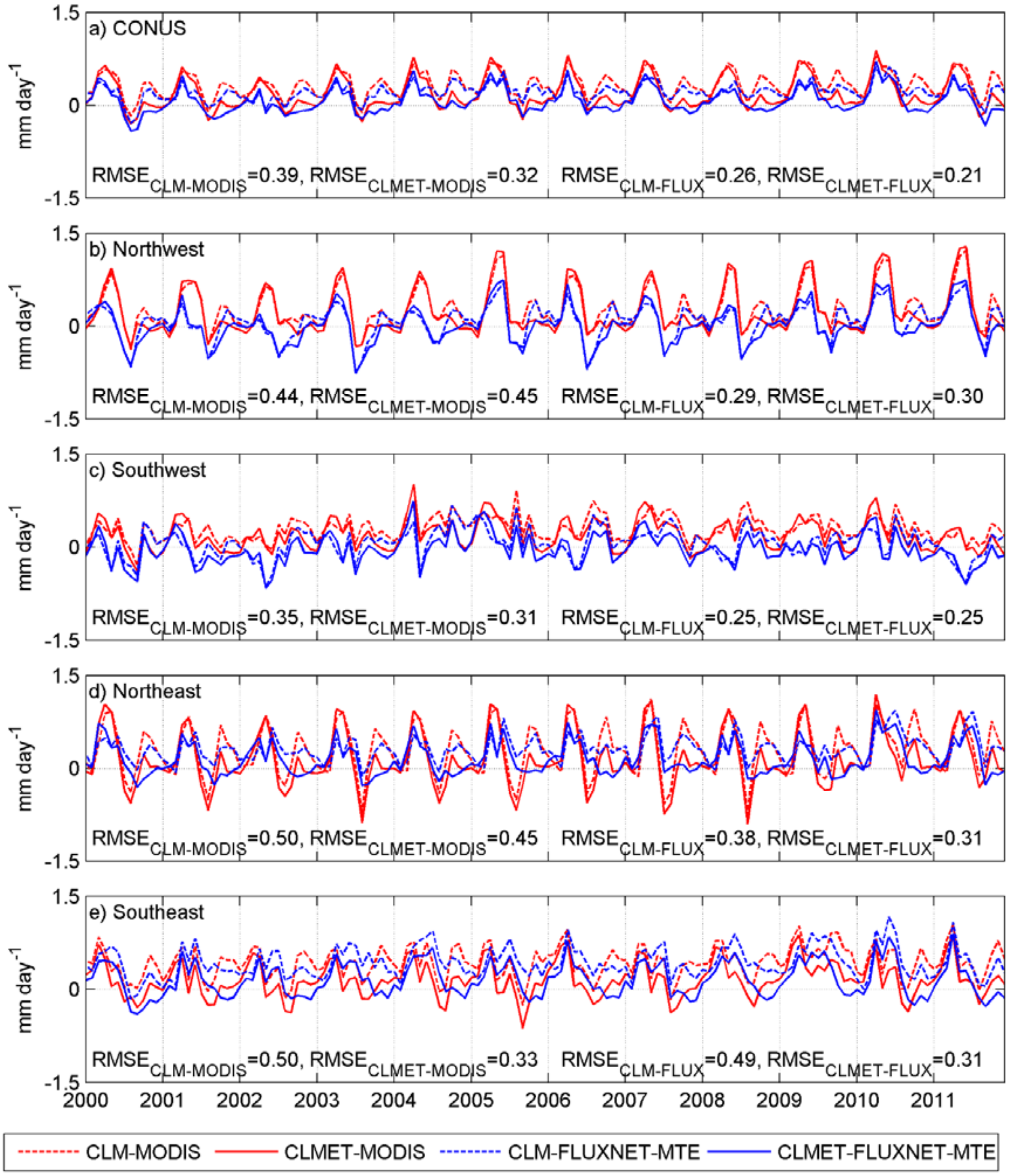

Figure 7 Time series of ET difference between model (CLM or CLMET) and reference data (MODIS or FLUXNET-MTE) over CONUS, Northwest, Southwest, Northeast, and Southeast during the period 2000-2011.

2. It is not clear to me how the statistics in Tables 1 to 4 are exactly calculated. This should be better documented in the manuscript. For instance, the temporal statistics in Table 2: are these calculated per pixel and subsequently averaged over the different study areas (CONUS, NW ...)? Or is the modelled evaporation first aggregated for the study area, and the statistics calculated on the aggregated values? In addition, next to the comparison against the FLUXNET-MTE product, I would also suggest to at least include a validation of the products against actual FLUXNET measurements. Although there are different issues with eddy-covariance measurements as well, a lot of data is

freely available, and these measurements are probably closer to the truth than any of the datasets currently used in the study.

Response:

1) On the calculation of the statistics in Tables 1 to 4:

We have added the following equations to the revised manuscript to show how the statistics is calculated.

$$Bias = \frac{1}{N}\sum_{i=1}^{i=N}\left(\overline{S_i} - \overline{R_i}\right)$$

$$Relative\ bias = \frac{1}{N}\sum_{i=1}^{i=N}\frac{\left(\overline{S_i} - \overline{R_i}\right)}{\overline{R_i}}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{i=N}\left(\overline{S_i} - \overline{R_i}\right)^2}{N}}$$

Where N is the total number of grid cells, and $\overline{S_i}$ ($\overline{R_i}$) are the temporal average of model simulated (reference) value for grid cell i, which is calculated as:

$$\overline{S_i} = \frac{1}{M}\sum_{j=1}^{j=M} S_{i,j}$$

$$\overline{R_i} = \frac{1}{M}\sum_{j=1}^{j=M} R_{i,j}$$

Where Si,j (Ri,j) is model simulated (reference) value on time j and at grid cell i, M is the total number of time series. The statistic RMSE is also used to validate models in reproducing temporal series where M becomes the total number of grid cells, and N becomes the total number of time series.

"In this study, the statistics Bias, Relative bias, and root mean square error (RMSE) are used to validate models in reproducing the spatial pattern against the reference dataset. They are defined as:

$$Bias = \frac{1}{N}\sum_{i=1}^{i=N}\left(\overline{S_i} - \overline{R_i}\right) \quad (1)$$

$$Relative\ bias = \frac{1}{N}\sum_{i=1}^{i=N}\frac{\left(\overline{S_i} - \overline{R_i}\right)}{\overline{R_i}} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{i=N}\left(\overline{S_i} - \overline{R_i}\right)^2}{N}} \quad (3)$$

Where N is the total number of grid cells, and $\overline{S_i}$ ($\overline{R_i}$) are the temporal average of model simulated (reference) value for grid cell i, which is calculated as:

$$\overline{S_i} = \frac{1}{M}\sum_{j=1}^{j=M} S_{i,j} \qquad (4)$$

$$\overline{R_i} = \frac{1}{M}\sum_{j=1}^{j=M} R_{i,j} \qquad (5)$$

Where Si,j (Ri,j) is model simulated (reference) value on time j and at grid cell i, M is the total number of time series. The statistic RMSE is also used to validate models in reproducing temporal series where M becomes the total number of grid cells, and N becomes the total number of time series." (the last two paragraphs of Section 2.2)

2) validation of the products against actual FLUXNET measurements

We have added validations against actual FLUXNET measurements at 16 stations. Please see the response to comment 1 for details.

3. I have the feeling that some issues of the method (e.g. the assumption of time invariant scaling factors or the use of monthly scaling factors) might be masked by the spatiotemporal scales at which the results are analyzed. For instance, why are only time series of the climatological cycle for the entire study area shown in Figure 6? It could be interesting to show some time series from individual pixels as well. Also, an analysis at shorter time scales might show some interesting results. E.g. why do the authors not show a time series of daily evaporation? The same holds for Figure 12: why are these time series not shown at daily time steps and on a pixel basis?

Response:

1) daily series of ET for individual pixels

We have included ET evaluation on daily and monthly scales at 16 pixels, and added figures to compare model simulations with in-situ observations. Please see the response to comment 1 for details.

2) daily series of soil moisture for individual pixels

It is difficult to determine which sites are suitable for validation from total 232 soil moisture observation sites. And the comparison between model simulations and site observations on the daily scale is consistent with the comparison on the monthly scale, as indicated by the comparison for ET. Therefore, we decide to still keep figures on the comparison at the state level (Figures 14 and 15).

4. P6-L116-117: Could the authors be more specific here about what is meant by spatial correlation? Observations from FLUXNET are essentially point measurements.
How are spatial correlations defined here?

Response: Parr et al. (2016) used FLUXNET-MTE (model tree ensemble) ET, which is a gridded ET product, to evaluate CLM4.5. We changed the description as follows:

The spatial correlation coefficients between the simulated annual ET and the FLUXNET-MTE (model tree ensemble) ET are as high as 0.93.

5. P5-L107: I think it should be mentioned here at what temporal resolution the model is applied. From the results in Table 2, I can guess the model is run at a daily resolution. If the latter is the case, I think it should also be justified why the scaling factors are calculated at the monthly time scale. Given that both the simulations and the GLEAM datasets are available at a daily resolution, the scaling factors could as well be calculated at the daily scale. Would this also work? Did the authors test the effect of applying daily scaling factors in the algorithm?

Response:

1) temporal resolution of model: the temporal resolution of model is one hour, which is typical for land surface models. We have added this information into Section 2.3 of the revised manuscript.

2) temporal resolution of scaling factor: This scaling factor characterize the relationship between model biases and ET climatology, and the fundamental assumption is that the nature of the model biases is time-invariant at the inter-annual and longer time scales. The monthly time scale is used here to account for its seasonality. To say that the nature of the model biases varies on a day-to-day time scale does not make physical sense, although technically it can be done. In fact we tested the performance of CLMET based on daily scaling factors. CLMET performance is not improved using daily scaling factors as compared with CLMET using monthly scaling factors.

6. P11-L244-245: Please revise this sentence: GLEAM data is not missing in this period, but is probably masked out in this study as the Northern regions of CONUS are typically covered with snow during these times of the year. GLEAM estimates of sublimation are available for these regions, but I guess they are not considered here (at P7-L140-141, it reads that only interception loss, transpiration and bare-soil evaporation are considered).

Response: we deleted the data records in some parts of west CONUS during the cold seasons by mistake, when GLEAM-derived ET is negative. This led to many missing values in annual ET map in Figure 4 of the original manuscript. We have corrected this mistake and updated Figure 4 (Figure 3 in the revised manuscript). The CONUS-averaged value from CLMET in the update version of annual ET is slightly better than the value in the previous version. We have also updated the table 1 to reflect this change.

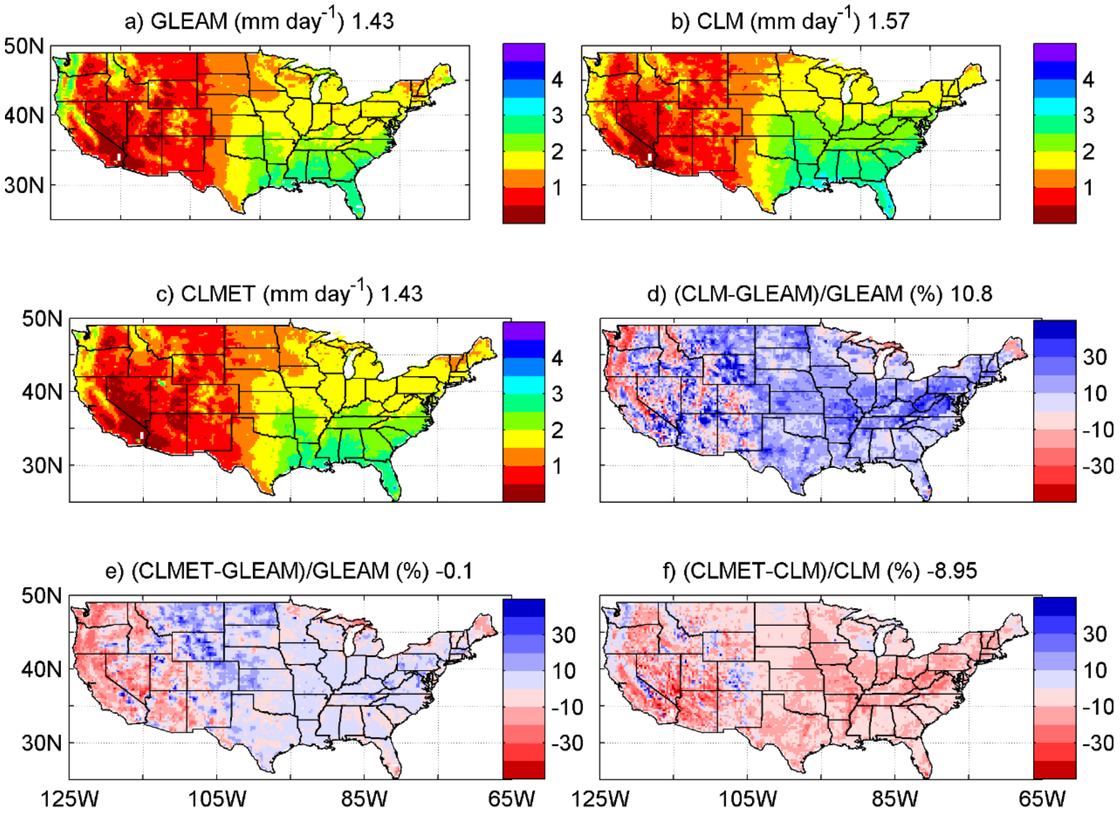
a) GLEAM (mm day^{-1}) 1.43
b) CLM (mm day^{-1}) 1.57
c) CLMET (mm day^{-1}) 1.43
d) (CLM-GLEAM)/GLEAM (%) 10.8
e) (CLMET-GLEAM)/GLEAM (%) -0.1
f) (CLMET-CLM)/CLM (%) -8.95
50N
40N
30N
125W
105W
85W
65W
4
3
2
1
30
10
-10
-30

Figure 3 Mean annual ET from a) GLEAM, b) CLM, and c) CLMET, and the relative differences between d) CLM and GLEAM, e) CLMET and GLEAM, and f) CLMET and CLM during 2000-2014. Numbers in titles are CONUS-averaged values.

7. P12-L261-262: If the term “significant” is used, it implies that a statistical test was applied to check this hypothesis. If this is the case, the test should be mentioned here.
Response: we changed to “substantially”.

8. Please note that the GLEAM datasets are no “observations” of evaporation. They are estimates of terrestrial evaporation, resulting from applying a simple conceptual model to observation-based datasets of different meteorological variables. GLEAM is kept as simple as possible to minimize the impact of the algorithms and maximize the impact of the meteorological observations on the estimates of evaporation. I would suggest to revise this throughout the manuscript.
Response: we have changed from “observations” to “estimations”.

TECHNICAL CORRECTIONS
1. Please use hyphens in “compound adjectives” such as “land-surface models” or “widely-used tools”.
Response: the expression of “land surface models” and “widely used tools” are widely used in literature.

2. I would suggest explaining all abbreviations upon their first use. E.g. P3-L68-69: SAC and VIC.

Response: Following reviewer's suggestion, we have spelled out SAC-SMA (Sacramento Soil Moisture Accounting) and VIC (Variable Infiltration Capacity) when they appeared for the first time in the revised manuscript.

"*The Mosaic and Sacramento Soil Moisture Accounting (SAC-SMA) models tend to overestimate ET, whereas the Noah and Variable Infiltration Capacity (VIC) models are likely to underestimate ET.*" (the last sentence of first paragraph, Section 1)

3. P5-L108: Given that no further details are provided in the paper regarding the land surface model used, I would suggest adding a reference here for the CLM model.
Response: Following reviewer's suggestion, we have added a reference about Community Land Model version 4.5 when the model is introduced in the revised manuscript.

Oleson, K. W. et al.: Technical Description of version 4.5 of the Community Land Model (CLM), NCAR Tech. Note, NCAR/TN-503+STR, doi:10.5065/D6RR1W7M, 2013.

4. P5-L111: Please define "PFT".
Response: Defined

5. P6-L124: I guess this should be section 2.2 instead of 2.3.
Response: Yes, it is 2.2. We have corrected it.

6. P7-L161: The fact that the GLEAM database has three subsets is not relevant here if you only use one.
Response: Following reviewer's suggestion, we have deleted the description of three subsets of GLEAM.

7. P28-Table1: Please correct "COUNS" in the caption. Please also check this at other places in the manuscript: e.g. P14-L315
Response: We have corrected them to "CONUS".

8. P34-Figure3: Please explain in the caption which areas are masked. I guess these are regions covered with snow?
Response: The GLEAM-derived dew and the CLM simulated dew is not consistent in some areas of northwest CONUS. If that happens, the scaling factors became negative, because ET is negative for one and positive for the other. We did not scale ET when the scaling factor is negative, and those areas are masked out in Figure 2. We have added an explanation about it.