Anonymous Referee #1 Received and published: 6 April 2021

Authors performed a multi-criteria evaluation of a continental scale integrated hydrologic model to identify sources of bias and error in hydrologic model predictions. While many continental scale hydrologic and land surface models have been developed and validated against observations, none of them have offered such a comprehensive evaluation using multiple point based and spatially distributed observations. Furthermore, many global scale models employ empirical formulations for simulating hydrologic processes to reduce computational time. Therefore, continental scale evaluation of a physically based model like PFCONUS is new. The study is comprehensive and very well written and organized. While model simulations for a longer time frame would be desired, I am aware of the scale of simulations and analysis done by the authors and only recommend a few minor comments below:

We thank the referee for their consideration, time, and thorough read of our manuscript. We have replied to your comments in blue text below. In some cases, we show text that has been added to or replaced sections of the original manuscript. The new additions or revised text are shown in green.

1) Add units of variables defined for each equation.

Dimensions have been added to physical equations (1) through (7).

2) Line 465- Add a label to Figure 3 to identify locations/extent of different river basins or label the basins in Figure 2.

Major basins have been labeled in Figure 2.

3) Could you please add NSE metric to Figure 4

Nash-Sutcliffe Efficiency was originally left out given some of the known problems with NSE as a performance criterion (Gupta et al., 2009). However, we recognize that the metric is widely used and that some readers and regional scale modelers would benefit from its addition. Given the manuscript's length, we have decided to add NSE in a supplemental text to the manuscript; we have also added the Kling-Gupta Efficiency (KGE, Gupta et al., 2009; Kling et al., 2012) in the supplemental information, as it addresses some of NSE's deficiencies and combines variability, bias, and correlation (Kling et al., 2012). The corresponding figure reporting NSE and KGE error metrics for PFCONUSv1 simulated daily streamflow is shown below (Figure S9).

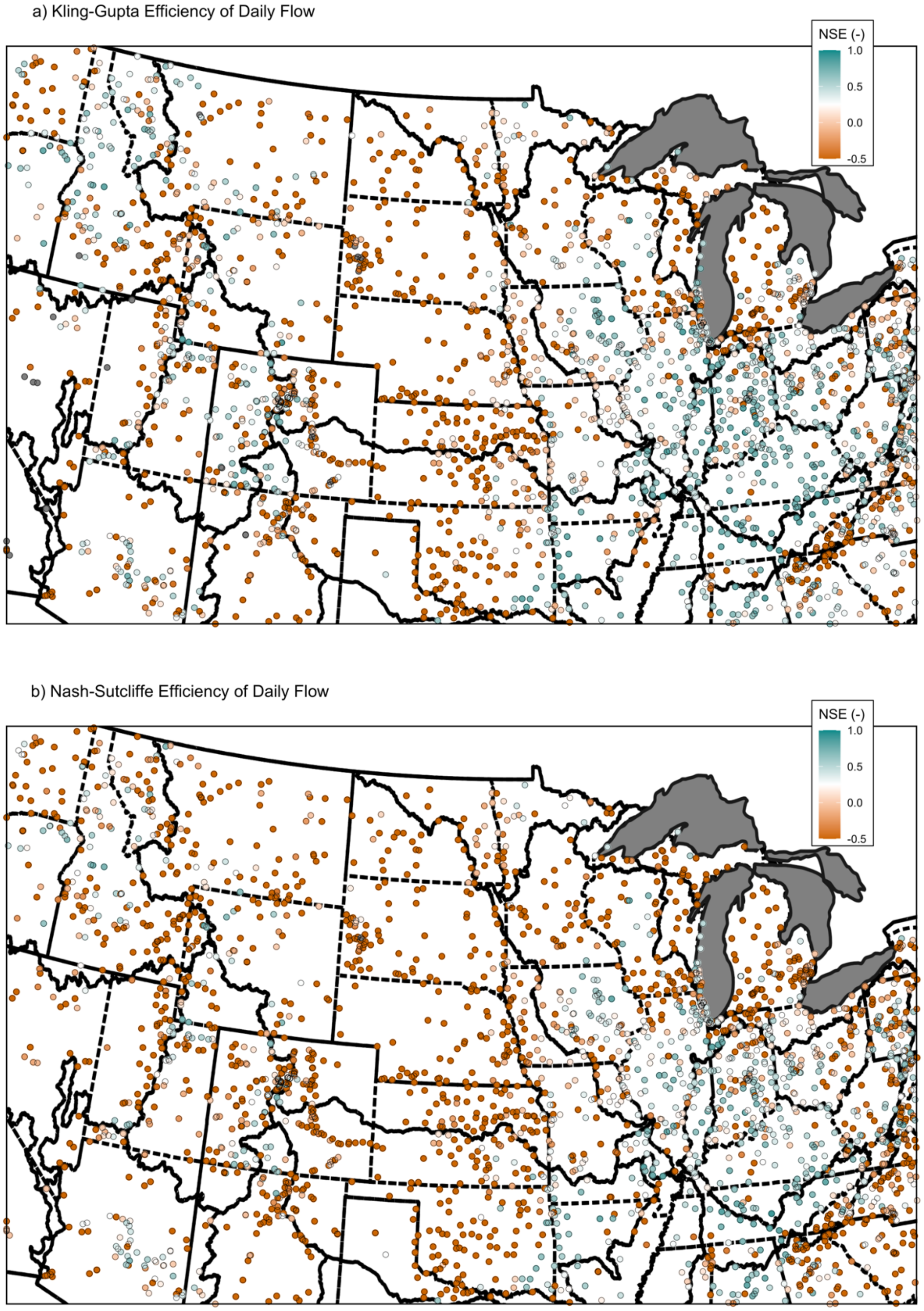

Figure S9: KGE (a) and NSE (b) evaluated at USGS stream gauges for PFCONUSv1 simulated daily streamflow.

4) Line 540 – Could you please compare performance of MOD16A2 and SSEBop against FLUXNET data using the same period as PFCONUSv1?

While evaluation of MOD16A2 and SSEBop algorithms with respect to FLUXNET observations is certainly valuable, we prefer to use the MODIS products at an aggregated temporal and spatial scale to evaluate PFCONUSv1 model results. As detailed in line 310 in the manuscript, there exist considerable uncertainties in point-scale MODIS values and high temporal resolutions associated with cloud cover and other limitations. We refer the reviewer to Velpuri et al. (2013) and Westerhoff (2015) for detailed evaluation of MODIS application at small spatial scales, and we feel that an adequate validation of the two MODIS algorithms against FLUXNET for the simulation period is outside the scope of this paper.

5) Line 561- Correct subplot number. In Figure 6, final row has the same subplot number as the previous row.

Thank you, this has been fixed.

6) Line 567- Correct Figure number. It should be Fig. 6

Thank you, fixed.

7) Line 690- Update figure number to Fig. 8d

Thank you, fixed.

8) For soil moisture comparison- Did you compare top layer simulated soil moisture with the ESA dataset?

Yes, only the top layer. This clarification has been added to the beginning of section 3.4.2, as well as to section 2.3.3.

9) Line 1325- ESACCI does not have mascon solution.

This typo has been fixed.

10) In Figure 9 – Does shaded region show standard deviation of spatially distributed soil moisture?

Yes, standard deviation was taken spatially across the major basin. This has been clarified in the figure caption for Figure 9.

11) Line 824- Change Fig 7g.h to Fig 6 g.h

Thank you, this has been fixed.

12) Line 841- Add “temperature”

Added.

13) Line 844-845- It is not clear. Please clarify.

The discussion here is referring to the fact that, on average and across all FLUXNET sites, ParFlow-CLM over (under) estimates low (high) rates of daily ET (Figure 13h) relative to observations. We have changed the text to clarify the point. This section now reads:

> "Overall, PFCONUSv1 under(over)-estimates relatively high (low) daily evapotranspiration rates (Figure 13h). For FLUXNET locations and days exhibiting ET rates over (under) 4 mm day^{-1}, mean daily bias is -1.2 mm (0.3 mm). Biases in NLDAS vapor pressure and wind speed could be a contributing factor. Lower vapor pressure deficits (0 to 20 Pa) and lower wind speeds (0 to 6 m s^{-1}) have an overall positive bias, which could explain PFCONUSv1 overpredicting low ET days. Similarly, we believe the bias on high-evapotranspiration days (ET > 4 mm day^{-1}), which PFCONUSv1 preferentially under-predicts, may be attributed to NLDAS under predicting wind speeds greater than ~10 m s^{-1}."

14) For discussion of uncertainty in meteorological forcing using ParFlow-CLM, I refer authors to Schreiner-McGraw and Ajami (2020, https://doi.org/10.1029/2020WR027639)

Thank you, we have included this reference in our discussion.

15) Figure 10- Add legend to subplots d and e

Legend showing the model simulation as blue lines/shading and the observed SNOTEL values as black lines/shading has been added to subplot e.

16) Figure 12 – Why the density of GHCND gauges are smaller in 12g-I compared to 12 j,k,l

Average daily temperature was taken directly from the GHCND network at meteorological stations with that metric available. Daily maximum and minimum temperatures were available at more GHCN stations than daily mean temperature; the GHCN network for mean daily temperature was simply less dense. We could in theory approximate mean temperature using the maximum and minimum values at more sites, but we believe this would defeat the purpose of using a quality-controlled network of observations to compare to our model results. The number of GHCN sites with available daily max/min and daily mean are detailed in the first paragraph of section 2.3.4.

References

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 370(1–2), 80–91.

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424–425, 264–277.

Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., & Verdin, J. P. (2013). A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment*, *139*, 35–49. https://doi.org/10.1016/j.rse.2013.07.013

Westerhoff, R. S. (2015). Using uncertainty of Penman and Penman-Monteith methods in combined satellite and ground-based evapotranspiration estimates. *Remote Sensing of Environment*, *169*, 102–112. https://doi.org/10.1016/j.rse.2015.07.021