1 Anonymous Referee #2

The authors provide a description of a publicly available dataset that they have developed for Germany. Their product will be useful for the scientific community. Aside from a few problematic oversights, the paper is generally well-written, with appropriate figures and references. In my opinion the paper will be suitable for publication after a minor revision.
Thank you for your helpful comments which are highly appreciated by us. The manuscript benefited from your suggested analyses and literature. We added the mentioned references to the revised manuscript and discuss them. Further, we address the questions: 1) Could the model performance of the 222 catchments be explained by any land surface or hydro-meteorological conditions?, and 2) How does the model estimate of ET compares with a remotely-sensed product? by additional analyses. In the following, we present the referee's comments as well as our point-by-point response to all of them.

1.1 Major

A major oversight of this paper is the lack of referencing a relevant paper that provides a similar dataset, at least in scope. The dataset of Newman et al. (2015) is also a 100-sample ensemble and needs to be cited here. The similarities and differences of the authors dataset with that of Newman et al. (2015) should be noted.
We discuss the difference in among these datasets in the introduction and the conclusions now. The mentioned references were added.
It is surprising that ET would have less uncertainty than streamflow since the latter is a more direct measurement. The authors only evaluate ET at 7 locations, while discharge is evaluated at over 200. It seems inconsistent to suggest that uncertainty across these two observations could be readily compared. Additional discussion is warranted here, including the scale mismatch between a 4×4 km^2 grid cell and a point observation.
The uncertainty of evapotranspiration and generated runoff is compared on the grid cell level, e.g., Figures 8 and 9. This comparison does not consider any observations. This analysis is based on the ensemble spread of the simulations at the 4×4 km^2 resolution. The model and model parameters are beforehand evaluated at point scale, i.e., 100×100 m^2, with observations at eddy covariance stations and at the 4×4 km^2 with discharge observations among others. For these evaluations we do not compare uncertainties of different variables as they would not be "readily comparable" because of the scale mismatch as you mentioned. So we fully agree with you in the argument that uncertainties from hydrologic variables at different resolutions are not comparable.
Further, the authors should comment more directly on why they did not evaluate the spatial patterns of their model against remotely-sensed ET and consider doing this evaluation.

Thank you for mentioning this point. We added a comparison of the ensemble mean of modeled evapotranspiration with MODIS evapotranspiration. We elaborated the results of this comparison in the revised manuscript and added Figure 1 to the manuscript).

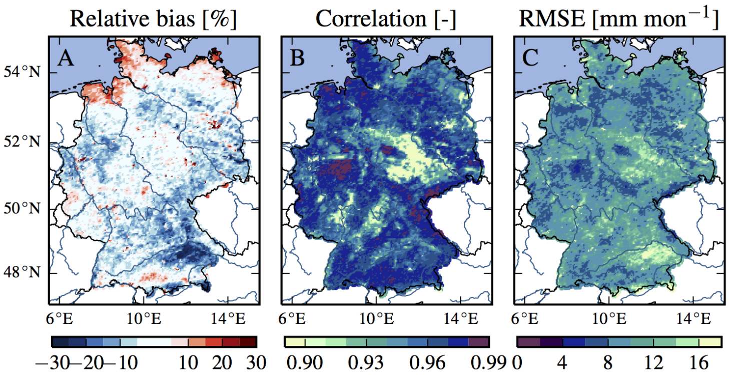

Figure 1: Comparison of monthly estimates of evapotranspiration from mHM and MODIS in the period 2001-2010. The ensemble is represented by the ensemble mean of 100 evapotranspiration estimates. The comparison is based on three statistical assessments: A) relative bias, B) Pearson correlation coefficient, and C) root mean squared error (RMSE). The respective units are given in brackets.

The validation watersheds range in size by nearly two orders of magnitude. If the model spatial resolution is the same for all, the authors should comment and hypothesize whether they see higher model performance in larger basins. Does performance increase monotonically with basin size?
We comment on this issue in the manuscript as “However, a tendency to perform better in large catchments basins is observed.” We also note that there is no clear (monotonic) relationship between basin area and NSE as can be observed from the Figure 2 below.
In Figure 4, climatic regime does not appear to be a good predictor of model performance, with some of the highest NSE scores distributed throughout the range of conditions. The authors should comment on what, if anything, will best predict model performance, to guide a potential user of the dataset.
Fortunately, we did not find any meteorological or morphological characteristics which explained why model performance is different for different catchments. This makes us confident that the retrieved parameter sets are representative for various climatic and physiographic conditions. We performed an analysis for identifying relations between land surface and hydro-climatic characteristics and model performance. Figure 3 was added in the revised manuscript and the findings are shortly discussed in section 4.2 of the revised manuscript.

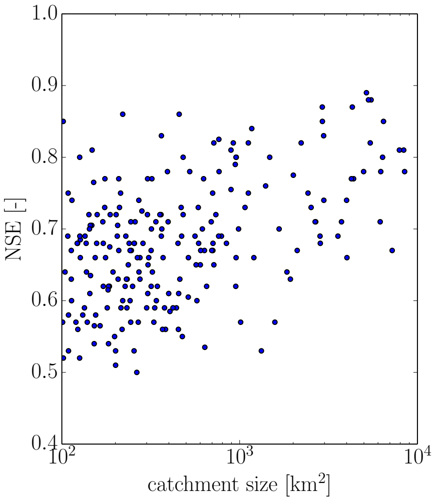

Figure 2: Relation of model performance and catchment area for the 222 basins.

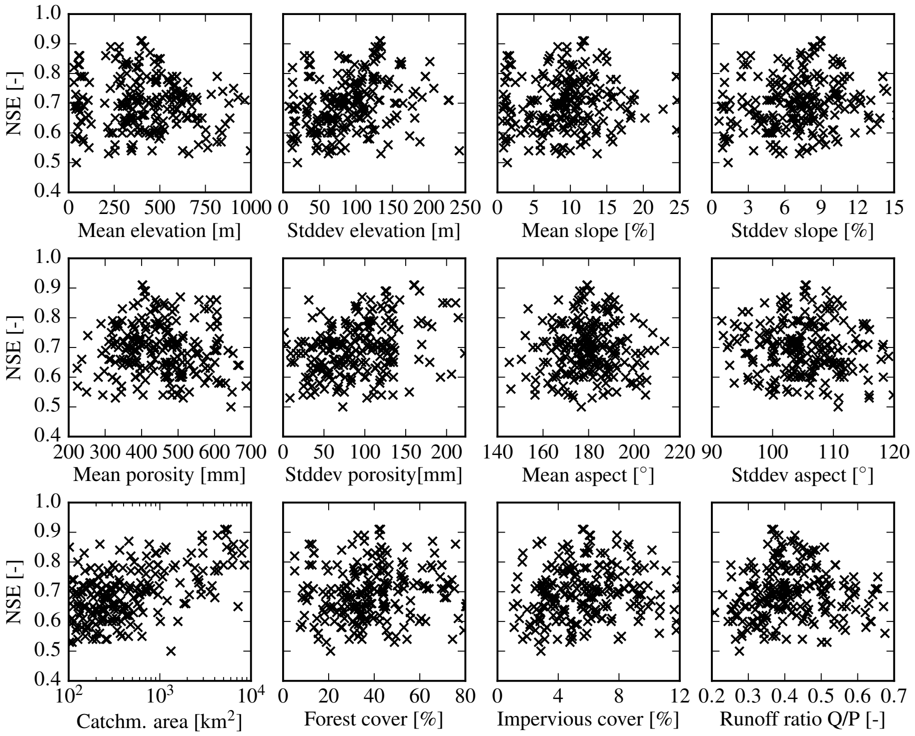

Figure 3: Relation between land surface and hydro-climatic conditions and model performance for the 222 river basins. The mean and standard deviation (stddev) of a characteristic for the single basins are based on the morphological input data at the 100×100 m^2 resolution. standard deviation.

1.2 Minor

P1L24: Grammar: have a footprints
Changed.
P1L24: 827 stations worldwide perhaps more apt to say "less than 1,000 locations worldwid", since there are other observational sources beyond fluxnet.
Changed.
P2L1: replace"reanalysis data" with "reanalysis products"and make this change elsewhere
Done.
P2L9: Maurer et al (2002) and Livneh et al. (2015) also cover a significant area in Canada (i.e. not just US, MX, and China).
Thanks for pointing out this fact. We changed the text "' accordingly.
P9: Here and elsewhere the use of the plural form of the word "performance" as"performances" is grammatically incorrect. Please correct this.
Changed.
References:

Newman, A. J., M. P. Clark, J. Craig, B. Nijssen, A. W. Wood, E. D. Gutmann, N. Mizukami, L. Brekke, and J. R. Arnold (2015). An observationally based gridded ensemble of precipitation and temperature data for the contiguous USA. J. Hydrometeorology, doi:10.1175/JHM-D-15-0026.1.

Livneh, B., Bohn, T. J., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., & Brekke, L. (2015). A spatially comprehensive, hydrometeorological data set for Mexico, the US, and Southern Canada 19502013. Scientific data, 2.