Supplement S7

October 25, 2021

1 Model comparison

A systematic model comparison was carried out based on the approach of Whitley et al.(2016), in order to provide more background for the model comparison in the main manuscript.. Two empirical benchmarks were created (emp1 and emp2), as initially proposed by Abramowitz (2012). Similar as in Whitley et al.(2016), emp1 is a linear relationship between GPP or ET and incoming radiation, and emp2 is a multi-linear regression between GPP or ET and radiation, air temperature and vapour pressure deficit. The empirical model for one specific site was created using the data from the other remaining sites.

Afterwards, a ranking of models was carried, similar to the PLUMBER methodology (Best et al. 2015). For this, the correlation coefficient, standard deviation, bias and normalized mean error were used. First, the models were ranked per statistical measure, and the average of these ranks was taken afterwards. This was done for each model in comparison with the two empirical benchmarks, as well as for all models together.

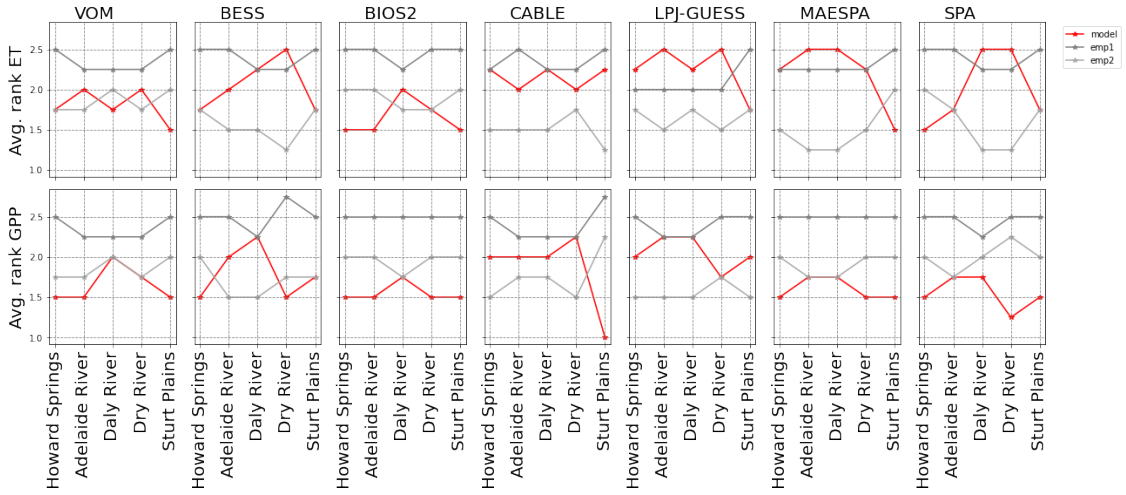

Figure S7.1. *The average rank per model in comparison with two empirical benchmarks (i.e. ranks go from 1-3) per site. The emperical benchmarks are based on Whitley et al.(2016), with emp1 a linear relationship between GPP or ET and incoming radiation, and emp2 is a multi-linear regression between GPP or ET and radiation, air temperature and vapour pressure deficit. First, the models and two empirical benchmarks were ranked for the correlation coefficient, standard deviation, bias and normalized mean error. Afterwards, the average of these ranks was taken. Model results are*

shown in red, empirical benchmark 1 and 2 are shown in gray and darkgray, respectively.

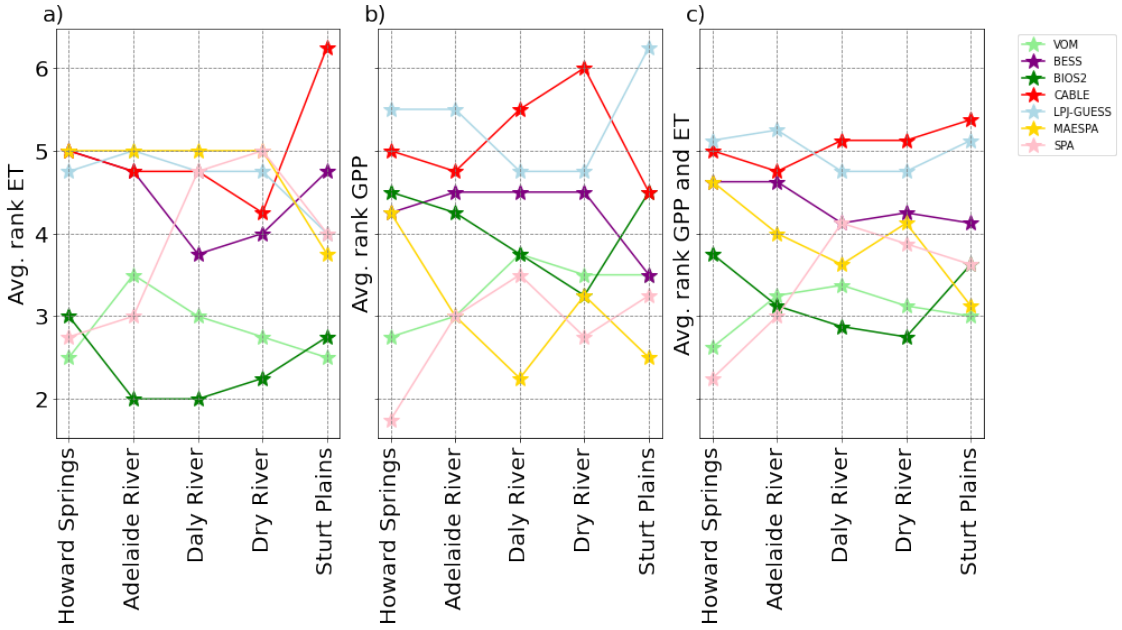

Figure S7.2. *The average rank for all models (i.e. ranks go from 1-7, with 1 considered the best rank) per site. First, the seven models were ranked for the correlation coefficient, standard deviation, bias and normalized mean error. Afterwards, the average of these ranks was taken for a) evapotranspiration performances, b) GPP performances and c) GPP and ET performances combined.*

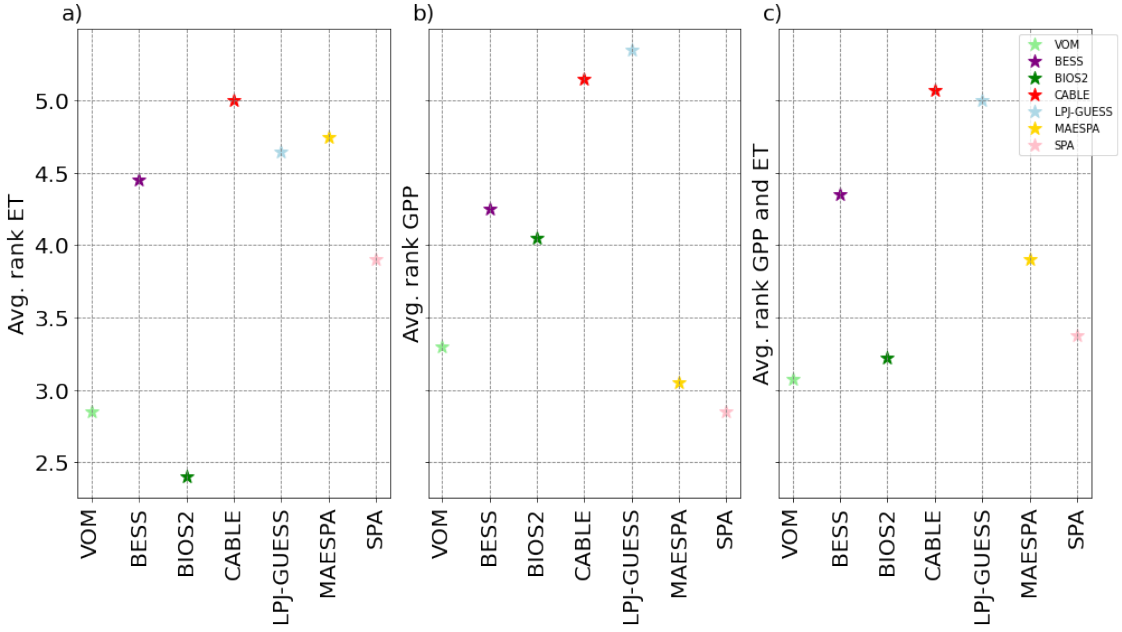

Figure S7.3. *The average rank for all models (i.e. ranks go from 1-7, with 1 considered the best rank), averaged over the different sites. First, the seven models were ranked for the correlation coefficient, standard deviation, bias and normalized mean error. Afterwards, the average of these ranks was taken for a) evapotranspiration performances, b) GPP performances and c) GPP and ET performances combined.*