SoMoBench: a quasi-global benchmark dataset for data-driven soil moisture modelling with in situ-derived targets

Sungmin O[1], **Rene Orth**[2], **and Wouter Dorigo**[3]

[1]Department of Advanced AI Engineering, Kangwon National University, Samcheok, Republic of Korea
[2]Faculty of Environment and Natural Resources, University of Freiburg, Freiburg im Breisgau, Germany
[3]Department of Geodesy and Geoinformation, TU Wien, Vienna, Austria
[*]corresponding author: Sungmin O (sungmino@kangwon.ac.kr)

Contents

Supplementary Figures **2**

Supplementary figures

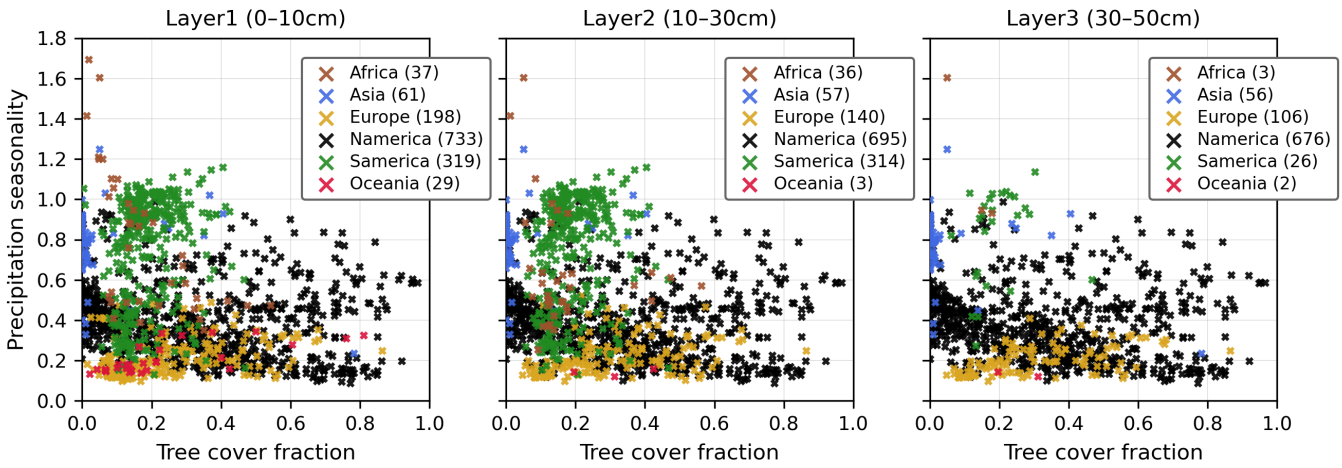


Figure S1. Environmental diversity of SoMoBench target grid cells. As in Fig. 2b of the main text, the distributions of target grid cells are shown for each soil layer, but here environmental diversity is defined by tree-cover fraction and precipitation seasonality. Colors denote continental regions, and numbers in parentheses indicate the number of grid cells.

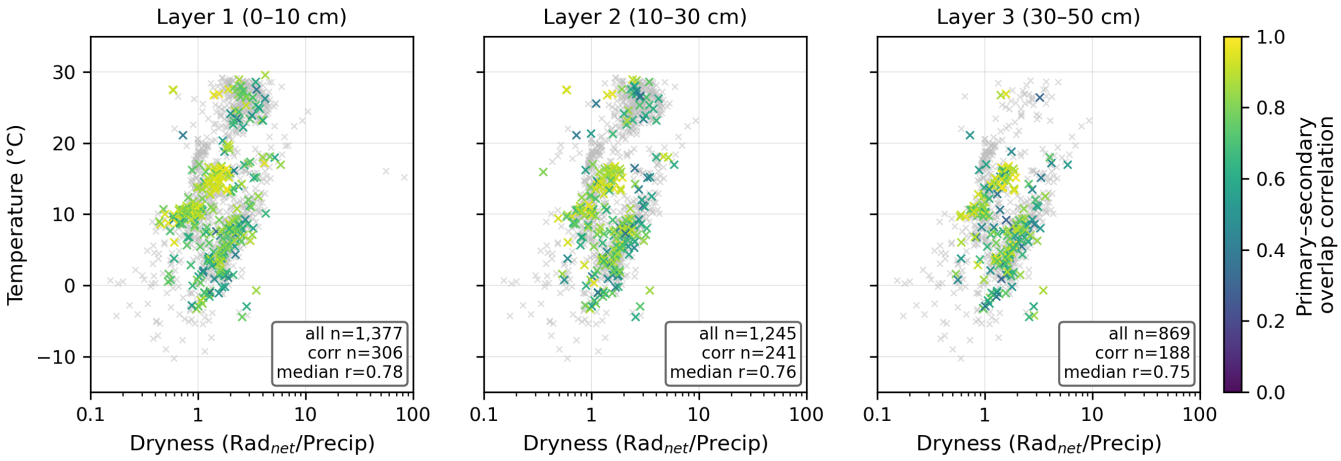


Figure S2. Hydroclimatic distribution of grid cells with contributing secondary stations. Grey crosses indicate all target grid cells in the same dryness–temperature space as in Fig. 2b of the main text. Coloured crosses indicate grid cells with at least one contributing secondary station and are coloured according to the primary–secondary overlap correlation. Insets report the number of grid cells and the median correlation for each soil layer.

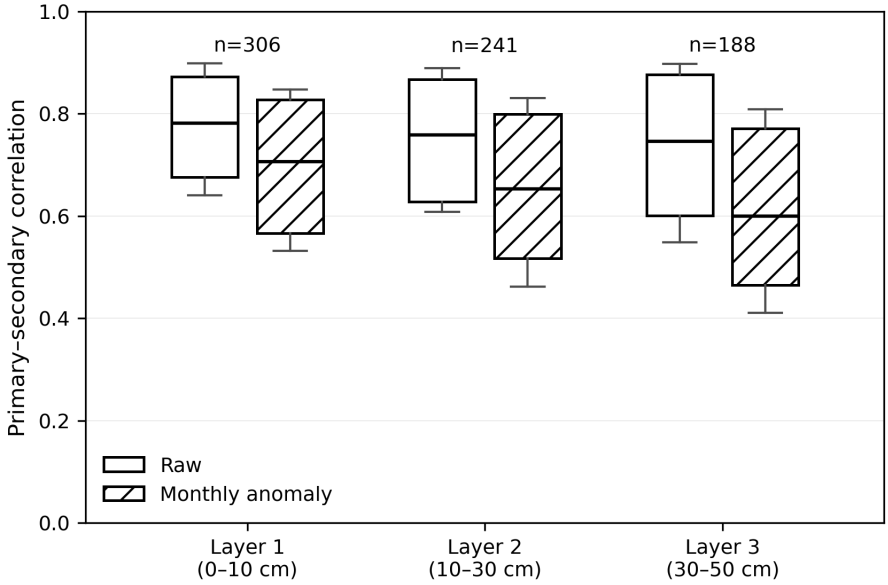


Figure S3. Primary-secondary overlap correlations for raw soil moisture and monthly anomalies. Raw correlations are the same as in Fig. 5b of the main text, and anomalies are calculated after removing each station's monthly climatology over the overlap period. Boxplot conventions are as in Fig. 5.

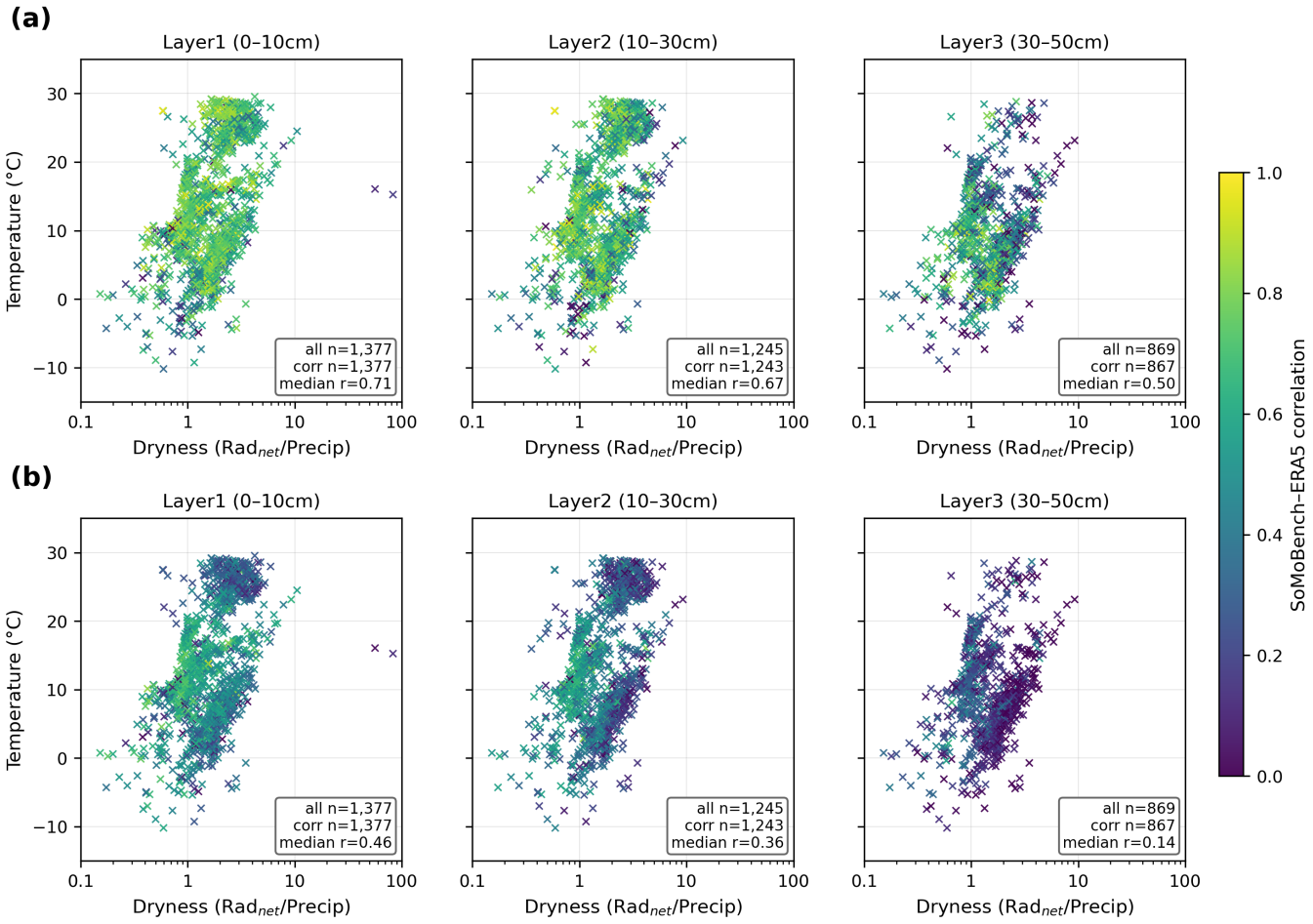


Figure S4. Hydroclimatic distribution of SoMoBench–ERA5 correlations. Same metrics as in Fig. 6b,c of the main text, but shown for all three layers across the dryness–temperature space: (a) daily soil moisture and (b) consecutive-day differences. Points are coloured by correlation, and insets report sample sizes and median correlations.