Thank you to the reviewer and editor for this further consideration of our manuscript. Below are the reviewer comments with our replies in bold.

I find the revised manuscript “Multi-scale assessment of a grassland productivity model“ to be much improved. I find the new introduction paragraph about “low-dimensional” models interesting and the expanded method section on estimating optimal parameter values (as well as the appendix) helpful.

Thank you.

• The authors emphasize in the responses to individual comments of both reviewers that the methods (e.g., model version, the input datasets, model evaluation) were carried out exactly as by Hufkens et al. 2016 – however, this point (and the rationale for why this was done) remains unclear in the manuscript. The method section in the manuscript should spell out that this work exactly re-created the methods used by Hufkens et al. 2016 (in terms of model formulation, input data, evaluation, etc.) but with more data, more sites, and more vegetation types. For instance, the section “2.3 Environmental Data” should mention that these datasets were chosen because these are the ones used by Hufkens et al. 2016 (and not because they are the most adequate), etc.

We’ve added text in the method and introduction sections to emphasize this point.

• I can understand that the authors chose to the Global Soil Data Task Group dataset because it re-creates the Hufkens et al. 2016 analysis – however, this reason needs to be explained in the manuscript (see above). In the response, the authors incorrectly state that none of the datasets suggested by reviewer 1 (see comment to original page 3 line 11) provide required variables (field capacity and wilting point). For instance, both NRCS soil data (gNATSGO, gSSURGO) and POLARIS provide water content at field capacity and wilting point (references in original comment of reviewer 1).

We apologize for not explaining further how the referenced datasets would not meet our needs. Specifically:

SoilGrids

According to the FAQ here (https://www.isric.org/explore/soilgrids/faq-soilgrids#Which_soil_properties_are_predicted_by_SoilGrids) there are no variables related to water holding capacity. The primary site (https://soilgrids.org/) also does not have any water related data.

WISE Soil Property Databases

According to the metadata here (https://www.isric.org/explore/wise-databases) and Batjes 2016 the only water variable is "Available water storage" which we assume is maximum saturated water content.

gNATSGO,gSSURGO

According to the dataset documentation here (https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053631) , specifically the "SSURGO Metadata – Table Column Descriptions Report" file, the only water variable is "Available Water Storage" at various depths. Which we assume to be maximum saturated water content.

POLARIS

Using the dataset description here:
http://hydrology.cee.duke.edu/POLARIS/PROPERTIES/v1.0/Readme

There are two water variables listed:

theta_s - saturated soil water content, m3/m3
theta_r - residual soil water content, m3/m3

Therefore we had to assume that field capacity and wilting point (commonly theta_fc and theta_wp) are not available. Unfortunately the POLARIS journal article (Chaney et al. 2016) does not go into more detail on available variables.

If field capacity and wilting point are indeed in these datasets then they should improve their documentation. Regardless we still would have used the original soil dataset for the reason above.

Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma*, *269*, pp.61-68.

Chaney, N.W., Wood, E.F., McBratney, A.B., Hempel, J.W., Nauman, T.W., Brungard, C.W. and Odgers, N.P., 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. *Geoderma*, *274*, pp.54-67.

• I do not follow the authors' response to reviewer 2 comment asking why "GPP data observed at fluxnet sites" were not used to evaluate PhenoGrass. The authors argue in their response that PhenoGrass predicts fCover and not GPP and that therefore GPP data cannot be used. However, the manuscript asserts that fCover is derived from Gcc by a scaling factor (supplement and page 2 line 23) and that "Gcc … is highly correlated … flux tower derived primary productivity (Yan et al., 2019; Toomey et al., 2015)". So, I do not understand why fCover cannot be compared against observed GPP from fluxnet sites.

We read the original comment as suggesting using flux tower derived GPP to fit the models, which is not possible with the current model form. Using flux data to further evaluate the model performance though, where Phenocam and flux towers are matched, is a reasonable suggestion and in fact done in Hufkins 2016. Unfortunately it would require considerable more time to incorporate this analysis, but we have included text suggesting this in the discussion.

• The revised version clarified that R2 was the coefficient of determination and named it unambiguously NSE. The last paragraph of the discussion section draws comparisons with Hufkens et al. 2016 (page 10 lines 6ff) "they had an average R2 of 0.71" – please clarify in the manuscript whether that is NSE or some other variant/interpretation of R2.

After reviewing the Hufkens 2016 supplement this was also the coefficient of determination and we changed in the text here to NSE to avoid any confusion.