**"Gap-Free Global Annual Soil Moisture: 15km Grids for 1991–2016" by Mario Guevara et al. Responses to anonymous reviewer #3**

REVIEWER COMMENT: General comments: While seeking a higher resolution global soil moisture product is certainly a laudable goal, I find the methods used in this paper lack a credible connection between ground measurements and the remotely sensed data. The authors use machine learning to generate regressions between multiple topographic variables thought to be related to soil moisture that are available at higher resolution with coarse resolution satellite data. It is difficult for the reader to discern if actual new information results from the downscaling because there is not a clear connection made between insitu soil moisture measurements, their physical connection to the chosen topographic variables used, and resultant satellite measurements.

AUTHOR RESPONSE: In a revised version we have improved the introduction and methods sections in a revised version to highlight why topographic terrain parameters are hydrologically meaningful. We would like to clarify that in situ observations are used for validating purposes only and were not used for developing (only testing) our downscaled soil moisture product. In this study, the main purpose was to compare if the fusion of satellite soil moisture records and elevation-derived terrain parameters (by the means of an empirical modeling approach, not physical) was useful to fill gaps of satellite estimates. The physical connection between soil moisture and terrain attributes is that these attributes regulate the overland flow and the potential solar radiation income (REF). Both processes are controlled by topography and therefore should show a direct influence on soil moisture patterns. We have also included more information supporting the use of these hydrologically meaningful terrain parameters for predicting soil moisture patterns, an emergent research opportunity on Geomorphometry. Finally, we clarified that our approach is not a process-based model, it is an empirical approach using machine learning and taking advantage of the large multivariate space of topographic parameters across the world (trained using ESA-CCI soil moisture) to predict soil moisture at 15 km resolution across the world. We highlight that this approach results in an improvement of the spatial resolution of soil moisture across the world and a better match with in situ soil moisture information (from the ISMN) when compared with the original ESA-CCI soil moisture product.

REVIEWER COMMENT: To do so, requires first demonstrating the rigor of the methodology over a much smaller and better measured area, such as the area of the International Soil Moisture Network.

AUTHOR RESPONSE: We fully agree with this comment. Our approach has already been tested across the contiguous United States (Guevara and Vargas, 2019). Thus, this study is an extension of that demonstrated methodology and now is applied to the global scale.

REVIEWER COMMENT: Extending the algorithm to areas that are hydrogeomorphically and climatically distinct from the existing soil moisture measurement networks cannot be result in credible data. I do not recommend publication of this work.

AUTHOR RESPONSE: We respectfully believe this is a misunderstanding from the reviewer. We clarified that or modeling approach does not use data from the existing soil moisture measurement networks to predict soil moisture across the world. Data from the ISMN is only used for validation purposes and was never used for training the model.

As explained in previous work (i.e., Gessler et al., 2009, Florinsky, 2012, MacMillan et al., 2016), terrain parameters (such as those we used for predicting soil moisture) both influence the accumulation of surface geological materials and reflect this spatial distribution. Terrain parameters such as the wetness index or the relative slope position both influence the flow and accumulation of moisture and reflect it. Terrain parameters both influence the spatial patterns of distribution of vegetation/land use and reflect these patterns. Our methods is based on geomorphometry which differs from a purely physical hydrological model (i.e., process-based models). That said the novelty of our study is that it takes the ESA-CCI and the terrain parameters in to a machine learning approach and finds relationships within the multivariate statistical space. The training data is the satellite soil moisture pixels in the latest version of the ESA-CCI, which are soil moisture values representative of a mosaic of ecological and environmental conditions (for a specific time) within an area (pixel) of around 27km of spatial resolution. The strength of a machine learning model as *kknn* is to find non-linear relationships in a complex multivariate space to predict patterns.

These patterns are dependent of training data (i.e., ESA-CCI soil moisture) and there are pixels with valid soil moisture data across dense vegetation areas or high latitudes that are used to train our model for soil moisture predictions (Fig_R3_1). We re-analyzed our datasets and updated soil moisture predictions including multiple sources of prediction factors (e.g., bioclimatic, soil type information) in order to test the reliability of our prediction framework.

The revised version of our manuscript included new data from other published soil moisture values to expand our validation dataset (Vargas et al., 2012, Saleska, et al., 2013). ). In addition, using in situ annual precipitation (Bond-Lamberty and Thomson, 2018), we report higher correlation between our soil moisture predictions and in-situ precipitation records, compared with the original ESA-CCI (e.g., from r=0.31 to r=0.38 in the tropics and r=0.40 to r=0.51 in temperate areas). We believe that this is good alternative comparison for validating and interpreting soil moisture predictions as previous studies have described the coupling between soil moisture and precipitation across multiples scales of available soil moisture and precipitation estimates (Koster, et al., 2004, McColl et al., 2017).

REVIEWER COMMENT: Major comment: Much more information is needed about the regression methods and parameter selection.

AUTHOR RESPONSE: In the revised version we improved the narrative of the kernel based machine learning algorithm. The main parameters of this method (*kknn*) are the kernel type (used to convert distances between neighbors points in the statistical space, to weights) and the k parameter (the number of neighbors used to calculate a weighted average in regression). These parameters are selected in our modeling approach automatically using repeated cross validation (Table 1 of submitted paper). The terrain parameters derived by the means of geomorphometry from the digital elevation model are another component of our modeling framework, used prediction factors (soil moisture covariates) in the *kknn* approach.

REVIEWER COMMENT: It would be helpful to discuss what each parameter is mathematically and why it can be useful in a soil moisture prediction context.

AUTHOR RESPONSE: We explain with more detail (in our revised manuscript) why the terrain parameters derived from elevation data using Geomorphometry are hydrologically meaningful. Detailed information on these terrain parameters is reported in our previous study across the conterminous United States (Guevara and Vargas 2019; see table: https://doi.org/10.1371/journal.pone.0219639.s007).

REVIEWER COMMENT: Can it be demonstrated over smaller areas that there is a valid argument for using these variables, some or all of them.

AUTHOR RESPONSE: Across the conterminous USA, we found an increase of nearly 25% (when compared with the original ESA-CCI soil moisture product) of accuracy validating our predictions against the North American Soil Moisture Database (https://doi.org/10.1371/journal.pone.0219639.g005).

REVIEWER COMMENT: Explain in detail the cross validation process used and why.

AUTHOR RESPONSE: Cross validation is a family of common re-sampling techniques that are used to analyze the sensitivity of models (in this case) to variations of available data for training purposes. Multiple models are generated using multiple proportions of available data and validating with subsets of information that are leaved out of these models. This process is repeated several times until capturing the magnitude of variance associated with the models based on the data subsets. In the revised version we included more information and references about cross validation strategies for re-sampling and bootstrapping prediction models.

REVIEWER COMMENT: It is unclear if there is a single model developed and applied to all years or separate models for each year. Are all topographic parameters used in the model in all years? What is their weighting?

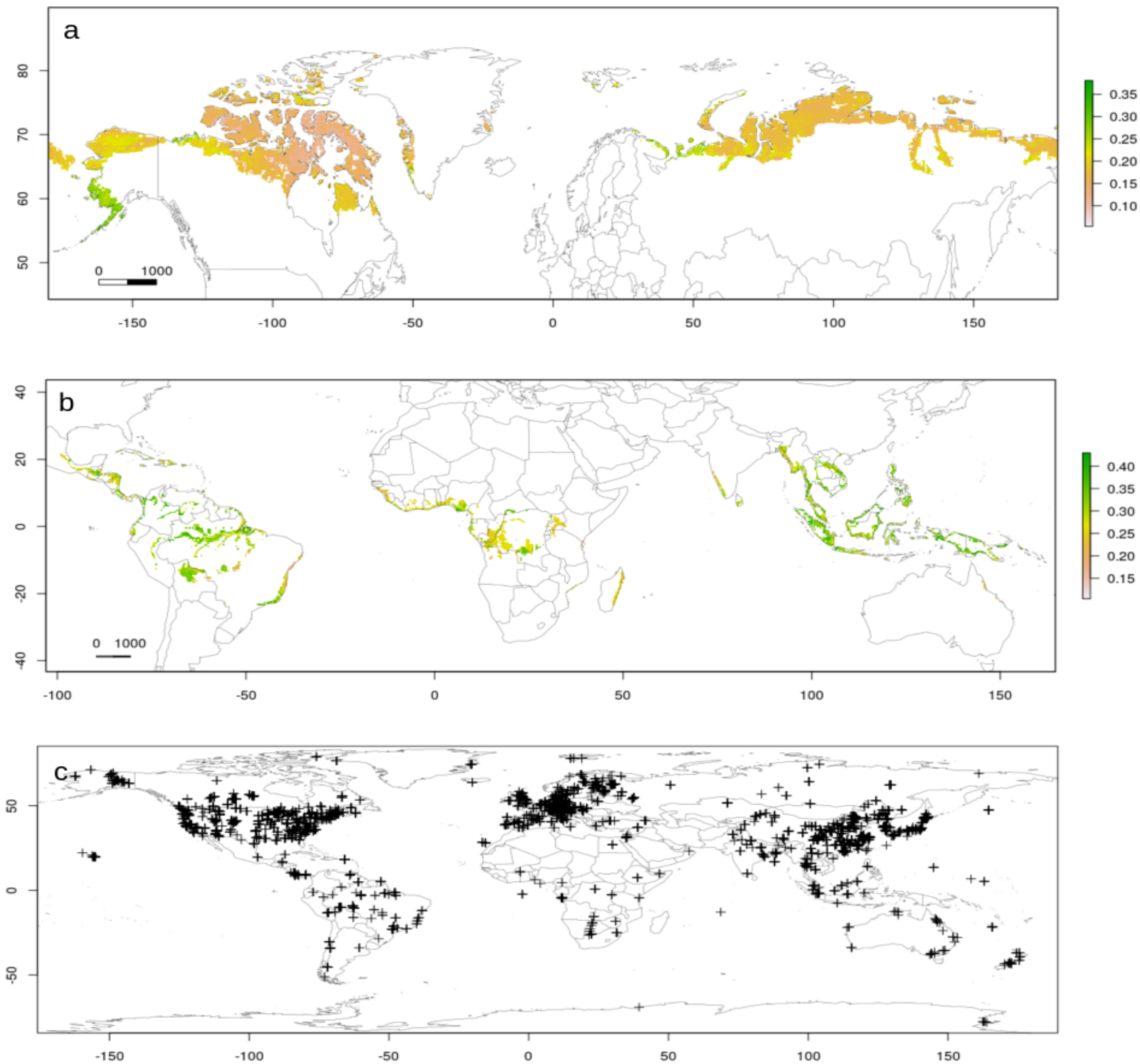
AUTHOR RESPONSE: We generated and cross validate a model for each year of available data. Table 1 (in the original manuscript) shows the results of cross validation and available data for each model/year. All topographic parameters are used on each model/year. During the cross validation, kknn uses a kernel form to convert distances in weights that are used for prediction. These distances are different from one place to another as the pattern recognition includes all the points and their k neighbors. Using the independent residuals of each model realization during the cross validation strategy, the optimal weights are selected using as indicators the root mean squared error and the correlation between observer and predicted (also included in Table 1 of the original manuscript). After all model realizations including all possible kernel combinations (e..g, "rectangular", "triangular", "epanechnikov","gaussian", "rank", "optimal") and increasing the number of neighbors (k) systematically (e.g., from 2 to 25, See Table 1) we find for each year the optimal weights maximizing the correlation between observed and predicted and minimizing the root mean squared error. In our reanalysis, we have included a novel variable importance analysis (by permutation) for kknn, that allows to identify which are the most important variables in the overall prediction using this kernel based algorithm.  We have clarified these statements in the revised version of the manuscript.

REVIEWER COMMENT: Major comment: Extending the results to areas without soil moisture measurements
(whether in situ or remotely sensed values) without any validation is not an improvement over current spatial data. It is highly suspect and could lead to inappropriate
applications using what is in effect non-data.

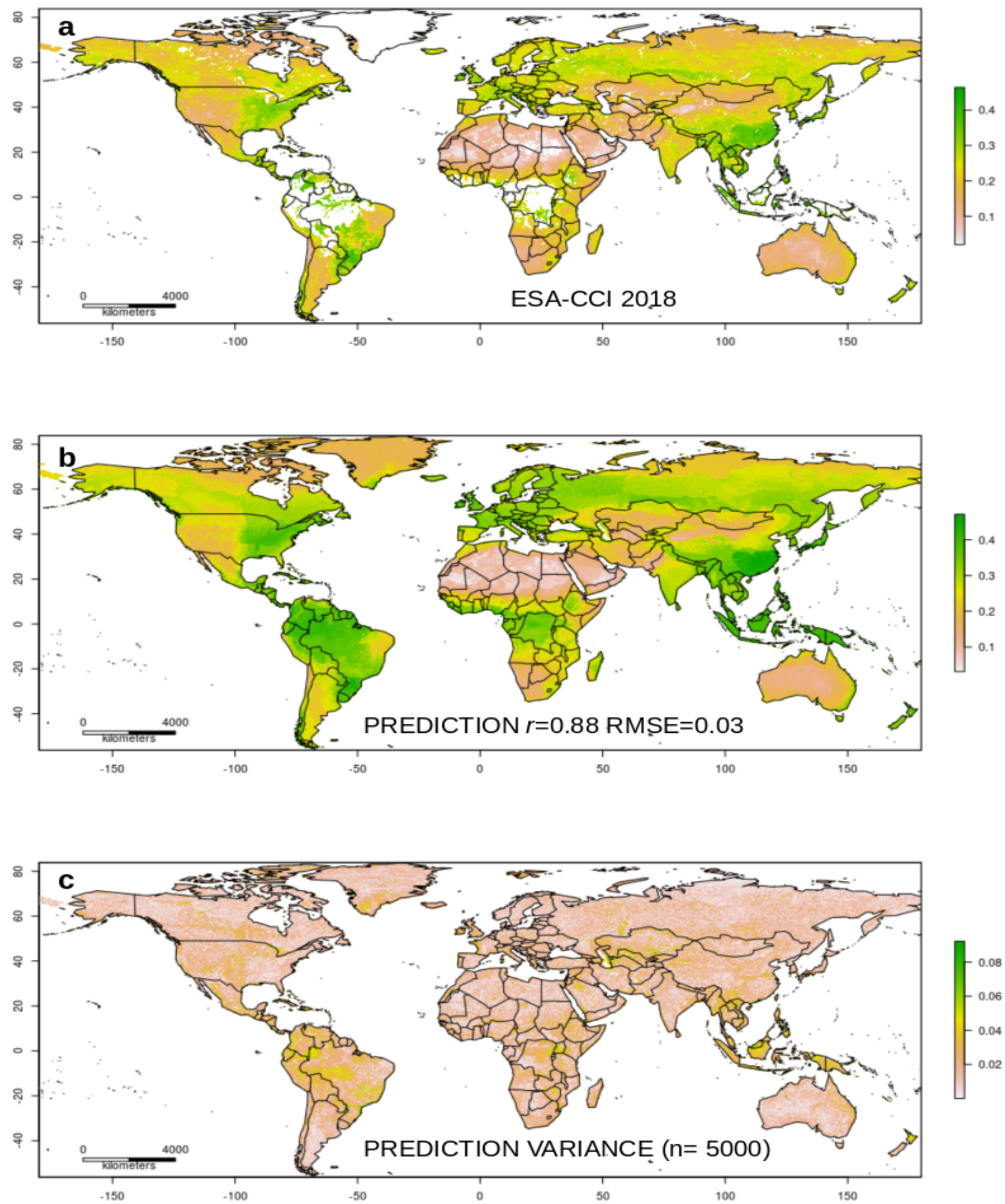
AUTHOR RESPONSE: We agree that any modeling output should be interpreted and used carefully as it is not a direct measurement. However, we performed a robust cross validation strategy including data available in the ESA-CCI representing all environmental conditions and provide information about data-model agreement. Then we validate with field data only at the places (pixels) containing in situ information in the International Soil Moisture Network and additional sites in tropical regions (9 more sites in the revised version of the manuscript). We argue that there are limited but representative satellite soil moisture records across these areas that can be used to train models for global soil moisture patterns (Fig_R3_1).

In the revised version we include more datasets (Fig_R3_1c) and more validation information to enrich the discussion about the reliability of our prediction framework for soil moisture (e.g., local precipitation measurements). We highlight that we are following a conceptual and data-driven framework assuming that comparing and testing multiple modeling approaches is still required to better understand the implications of soil moisture on ecological patterns across areas where no soil moisture information is available. Thus, our results provide evidence that our soil moisture product is useful to predict soil moisture patterns in places with low availability of satellite soil moisture estimates.

Finally, in the revised version we now include uncertainty estimates to complement this study and inform users and applications about model performance across the world (Fig_R3_2).




Fig_R3_1 Available information in the ESA-CCI for the year 2018 in the latest version product (4.5). We show the available information across Polar environments (a) and across dense vegetated areas and tropical rain forests (b). We also show a recent database with in situ information of climate variables and ecological properties that we used to support the reliability of our prediction framework in a revised version of our paper (c).




Fig_R3_2 Mean soil moisture for the year 2018 from the ESA-CCI (a). Spatial prediction of soil moisture for the year 2018 across 15km grids (b) and prediction variances based on bootstrapping the spatial prediction model.

REVIEWER COMMENT: Moderate comment: I found the paper very repetitive and in need of detailed editing.

AUTHOR RESPONSE: We have revised the narrative and overall organization in the revised manuscript.

## References


Bond-Lamberty, B.P., and A.M. Thomson. A Global Database of Soil Respiration Data, Version 4.0. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1578, 2018

Florinsky, I. V.: Influence of Topography on Soil Properties, Digital Terrain Analysis in Soil Science and Geology, 145–149, doi:10.1016/b978-0-12-385036-2.00008-0, 2012.

Gessler, P., et al., Chapter 28 The Future of Geomorphometry, Developments in Soil Science, 637–652, doi:10.1016/s0166-2481(08)00028-7, 2009.

Guevara, M. and Vargas, R.: Downscaling satellite soil moisture using geomorphometry and machine learning, edited by B. Poulter, PLOS ONE, 14(9), e0219639, doi:10.1371/journal.pone.0219639, 2019.

Koster, R. D.: Regions of Strong Coupling Between Soil Moisture and Precipitation, Science, 305(5687), 1138–1140, doi:10.1126/science.1100217, 2004.

MacMillan et al., A hydrologically explicit, spatially exact, classification of landforms for Canada at 1:500,000 scale., EGUGA, EPSC2016-13103 https://ui.adsabs.harvard.edu/abs/2016EGUGA..1813103M/abstract (Accessed 23 December 2019), 2016.

McColl, K. A., Alemohammad, S. H., Akbar, R., Konings, A. G., Yueh, S. and Entekhabi, D.: The global distribution and dynamics of surface soil moisture, Nature Geoscience, 10(2), 100–104, doi:10.1038/ngeo2868, 2017.

Saleska, S.R., H.R. da Rocha, A.R. Huete, A.D. Nobre, P. Artaxo, and Y.E. Shimabukuro. 2013. LBA-ECO CD-32 Flux Tower Network Data Compilation, Brazilian Amazon: 1999-2006. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA http://dx.doi.org/10.3334/ORNLDAAC/1174

Vargas, R.: How a hurricane disturbance influences extreme CO2fluxes and variance in a tropical forest, Environmental Research Letters, 7(3), 035704, doi:10.1088/1748-9326/7/3/035704, 2012.