Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-76-RC1, 2022

Comment on hess-2022-76

Anonymous Referee #1

Referee comment on "A robust gap-filling approach for European Space Agency Climate Change Initiative (ESA CCI) soil moisture integrating satellite observations, model-driven knowledge, and spatiotemporal machine learning" by Kai Liu et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-76-RC1, 2022

OVERALL IMPRESSION

The study presents a method for gap-filling ESA CCI soil moisture data. For filling the gaps, the approach utilises information from the spatiotemporal domain around the missing value as well as from other explanatory variables. This is a timely contribution to the ever-growing field of gap-filling and data fusion in Earth system science. The study benchmarks the method over China, which covers a large variety of different climate zones and topography, making it a suitable test bed. Particularly with the severity of missing values in microwave remote sensing soil moisture retrievals, it is vital to use as much information as possible, both from the spatial and temporal domain as well as from other available observations. Therefore, the proposed method seems promising to be able to do the task. Furthermore, comparing the gap filled values to in-situ observations gives an independent insight into gap filling performance.

However, the poor use of language and grammar hinder understanding of the methods and results in many cases. Additionally, the structure of the results and methods sections are a bit unclear and seem arbitrary, such that following the storyline of the paper is difficult. Finally, the purpose, aim and implementation of some of the methodological choices and evaluations in the result section are unclear and should either be replaced or clarified.

I therefore think the paper requires major revisions. Please see below my general and specific comments.

GENERAL COMMENTS

The study needs restructuring especially in the methods and results section. It is hard to follow the methods and results section, since they are structured very differently. Please align Figure 2 with the structure of Section 3, such that it is easy for the reader to follow which part of the model is described. For example, the subsections could have the same header names as the boxes in Figure 2. Also, it is not clear why only parts of Figure 2 are described in the subsections. Furthermore within the results, sometimes two different results are explained in one subsection (e.g. Section 4.2. compares to in-situ measurements and to the "cross-validation", Section 4.3 compares to literature and to other employed methods). I suggest making individual sections for individual results, or clarify why the results are thrown together in one Section like this.

In general within the result section, the results are described with generic terms like "good coindicence", "capacity for reconstruction", "high accuracies", "strong variation delineating capacity". Whilst those all are incorrect English terms, they also do not go into detail or describe the results well. While it is important to give statistical evidence of the algorithm performance, it is of equal importance to discuss the difference between the statistical measures and possible reasons for it, as well as evaluating the physical plausibility and coherency of the gap filled values. It would highly benefit the study if the results were described more concisely, more descriptive words would be used, and the results would be discussed, taken into context and described better. A good example of a good explanation of results is for example found in L433-436. Please more of those. Furthermore, it would be beneficial to add at least one plot that evaluates the physical plausibility of the gap-filled values for a possible application of the gap filled dataset. As an example, since soil moisture values are often important when analysing droughts, it would be interesting to see whether the gap-filling method is able to not only work in the mean of the values, which all the statistical metrics aim at, but also see whether the extreme values are gap-filled well. This could be done by showing some maps of a known drought event over China, possibly compare with the in-situ measurements and the original, gappy CCI SM data.

Throughout the script, there are many spelling and grammar mistakes, some of which

significantly hamper understanding of particular methodological decisions or results. It is vital that these mistakes, some but not all of which are listed below, should be removed from the script. A thorough cleanup of the language and individual sentences is necessary. Many statements could also be reduced, whilst keeping the meaning, to a minimum of words, effectively improving the structure and readability of the script and at the same time reducing its size.

It is a reasonable and thoughtful decision to investigate the importance of the different selected variables for the gap-filling process. However, I am not yet fully sure whether the selected method is the best in this context and whether it is applied correctly. Firstly, the mentioned model produces a significance score that I cannot assess, since the equation is not provided. It would be vital to add this, since only then the meaning of the significance (e.g. in Fig 3a) can be understood. Furthermore, the cited literature uses this method only for exploratory variable importance in hydrological case studies and not to limit variables that feed a gap-filling model. In the latter case I suggest another method that is well-known and often used in the machine learning context for feature selection prior to model employment could be used. These methods include the feature importance assessment of the Random Forest model, a permutation feature importance assessment, introducing a regularised regression that puts weights on the input features, or evaluating using the SHAP value. Please provide in your answer to this review a discussion on which method you settled and why you chose this one. Furthermore, the used method is a linear regression, which screens for important variables only in linear relationships to soil moisture. However, the relationship between soil moisture and the provided explanatory variables is in many cases non-linear, therefore there is a chance that important dependencies are missed if only for a linear relationship is screened. Additionally, the missing regularisation in the linear regression makes it prone to overfitting. Finally, it is unclear to me why a variable selection / importance assessment is conducted twice, in Section 3.1 and in Section 4.4. I suggest only making one of those assessments, e.g. by removing Section 4.4. and potentially replacing the assessment in Section 3.1 with a more robust method, see my suggestions above.

Why do you use a combination of scaling algorithms in Section 3.1.2? Since you never compare ESA and ERA soil moisture values in the results, you could just compute standardised anomalies of the ESA values, run the gap filling model and then add mean and standard deviation again, to convert back to physical values? Please explain, in case I have missed why this step is necessary, or just use standardised anomalies for simplification.

In result Section 4.1 you compare the original daily CCI SM and the corresponding gap-filled dataset for the year 2009. You argue that the model performance has merit based on the fact that the two datasets show similar characteristics in Figure 6. This however is not an argument that is relevant in gap-filling. Since the missing values in the original CCI SM data are missing not (completely) at random, but are missing systematically where vegetation cover is high or the soil is frozen, this dataset is biased towards the underlying, unobservable gap-free "truth". Comparing the original, gappy data with the gap-filled that is supposed to produce biases, i.e. change the statistical moments of the data because it is missing not at random (see e.g. Rubin et al, 1976, Little et al 2014, Van Buuren et al 2018, Bessenbacher et al 2022). If these two datasets are compared, it should be aimed towards physical plausibility and coherency, and not about whether they are statistically similar.

It is unclear why a cross-validation is performed on presumably the year 2009. L335 only cites generically "model performance" as a reason. A cross-validation is usually performed to find the corresponding parameter values of the model (here the Random Forest), but this procedure is already described in Section 3.2 and refers to the years 2003-2008 which are never shown in the result section. Which parameters are defined with the cross-validation as described in L334-335? Which values are Figures 10 and 11 comparing to? The purpose and aim of this part of the results section is unclear to me.

Section 4.5: It is unclear why suddenly focus regions are introduced and used, and why this analysis cannot be conducted on the whole area (all China). Also, it is unclear why the focus regions are so different in size. This makes them harder to compare, as the wet region has less datapoints to compare and covers a much less diverse climate zone (compare Figure 1). Please clarify why these decisions are necessary or make the analysis on the whole of China. Furthermore within this section, please verify that none of the SM data from ESA, GLEAM and Noah is used for ERA. If that would be the case, the datasets are not independent, and this could for example explain that they are more similar. Similarly, check that neither MODIS or GLDAS are used for Noah or ERA runs.

Since this gap-filling method could be an important contribution to the problem of missing values in soil moisture remote sensing observation and Earth observations in general, it would be beneficial if the code of the method could be published, preferably on a platform that enables easy use for interested users and a versioning system (e.g. Github). Also for the purpose of this review, if possible, it would be interesting to have a look at the code to understand better what is going on and how exactly this is implemented.

SPECIFIC COMMENTS + TECHNICAL CORRECTIONS

L20 “Compared to that…” I don’t understand this sentence. Please clarify

l31: “SM has been declared” please add citation

L44 there is some literature on the shortcomings of soil moisture assimilation into reanalysis, see e.g. Dorigo et al 2017

L62: “some studies” but only one study is cited.

L74 and all other occurrences throughout the text: the word “delineating” is used in an unusual way. I suggest replacing it with depict/represent/show/ or similar in all occurrences

L89 and all other occurrences throughout the text: using “the” in front of an previously unmentioned fact is confusing. Replace with “a” if not referring to a specific one.

L95 consider citing Bessenbacher et al 2022

Fig 1: please increase resolution. Explain acronym DEM at first occurrence.

L119 “mainly”. What is not included in this list?

Table 1: remove lines below “model analysis” in first column. Left-aligning columns could improve readability

L 121: mention already here what the difference between “model establishment” and “model analysis” is. Is one the features used to run the model, and the other the evaluation? Not clear.

L130: data is a plural word. Always “data are”

L142: to better understand the negative correlation of TPI and CCI SM in Figure 3 please add a short sentence explaining what a high / low TPI mean.

L159: this is a good example of how the text can easily be shortened without losing understanding. "Precipitation, air temperature, … are obtained from the Chinese …". Please look for those sentences elsewhere as well to clear up text.

L160: since this is dataset from ground stations, but you state it is gridded, how was the gridding procedure performed? Is there literature that you could cite?

Table 2: Since you never discuss results at the individual stations of the WATER and CERN stations, I don't think it is necessary to add a table here naming them all. Consider moving to Supplementary Material or adjust such that table only includes networks and not individual stations.

L184 please add citation for this claim.

L194 "a vector the sample number of which is decided by" no correct English sentence.

L197ff: I don’t understand how steps (i) through (iii) are related to the four boxes in Figure 2. Adjust Figure 2 such that it has the same structure as the text, or vice versa.

Fig 2: Explain colours and frame shapes. Explain which criterion is to be met in “data judging”. Explain which explanatory variables are taken from “dataset preparing” to “dataset judging”. Rename judging.

L214: “plenty” as in all possible combinations? Are you stepwise removing or adding variables, or are you trying all different combinations? Please clarify.

L217: please define “importance criterion”

L223: what happens in the case that gaps are present in the variables?

L224: are slope, lat, Lon, aspect and wind removed or not? They are not visible in Figure 3b,c but their removal is not mentioned in the text.

Eq4: what does subscript p1 mean? What does the “.” And the “‘ “mean in the second

equation? Isn't mu(SM_c1) the same as mu(SM_ESA (t_av))? If so, please simplify equations.

L254f: please define the difference between "traditional regression-based methods" and "machine learning approaches", since both are regressions, machine learning models do not inherently come with uncertainty estimation (for example, a Random Forest model does not have intrinsically uncertainty estimation and you don't have it in this study) and it is not less likely to overfit with machine learning methods.

L260 "is feasible to add layer categories" please clarify.

L293: please clarify how missing values in the explanatory variables are treated within this algorithm.

Fig 5, caption: please define sw and nd for better readability

L308: please clarify "neglectful variables"

L311: please introduce acronym GWR before first mention.

L319: typo “her”

Section 3.4: Is this just a linear regression, applied to the same time window approach as the Random Forest, applied on the residuals from the Random Forest model? Please clarify

L322: Please show a plot that shows the relative contributions of the GWR interpolation and the Gaussian Filter smoothing in the reply to this review, as to see that the influence of the latter on the results is smaller than the one of the former.

L354, 355: “heavy missing issues”, “relative minor conformity” please correct English and clarify meaning.

L356: “consistent pattern” please clarify. Are you arguing they are similar or they have consistent biases? If yes, which? Please describe the results more.

Fig7,8: Please put the corresponding days next to each other to simplify comparison

L388: I disagree that the values are close to the 1:1 line, but I also think that this is hard to achieve given the spatial gap between point measurements and gridded measurements. Would discuss here shortly.

L391: please clarify sentence “in general..”

L399: NSE is not introduced in Section 3.5. Please add.

L402: please clarify sentence “in general..”

Fig9g: please add the fraction of missing values for each day (e.g. similar to precipitation bar plots) such that it can be evaluated how the gap-filling performs with little or many missing values.

L423 please explain mechanism better

L431: again please refrain from simply describing the results as “good” without further analysis (see also general comment above)

L442 typo “in suit”

L442 “satisfied performance” please correct English

L446 “severe missing issues” please correct English

L453 “future compared” please correct English

L465 since the 9% & 19% increase in accuracy stemming from residual calibration and the spatiotemporal domains, respectively, is an important result that you mention in the conclusions and in the abstract, it should be more clear in Figure 12. Maybe add a figure with accuracy change per change in the method?

Figure 12,13, 14: R**2 doesn’t have a unit (Accuracy, cm**3/cm**3). Please clarify. For example, add an optimal value to each score (0 for RMSE, 1 for R2) and sort the diagram after scores, not after methods, such that the scores can be directly compared.

L514 please clarify where this result is visible. If mentioning percentages of chance in the text, they should be visible in the graphs directly and not from comparing different bar plots visually.

L532 “has a long sequence” please correct English

L534 “more than 90%…” this is not clearly visible in Figure 15

L535 “comparable accuracy”. Some metrics are better, some are worse. Discuss differences and possible reasons!

L545 Please explain the mechanism better.

Table 5: please reorder such that the values are immediately comparable, e.g. the R2 columns next to each other etc

Fig 15: use different colorscale for last column to improve readability. For example, blue to red, for “wetter” to “drier”

Fig 16: one plot per climate zone, not one plot per dataset, such that they are visibly comparable. Also, disaggregate into seasonality and interannual variability to further analyse if both characteristics of the soil moisture dataset are reproduced in the gap-filled version.

L576 “study presents” please correct English

L585 “especially for areas with large swath gaps” this is not shown in the results but would be very interesting

L594 “manifest” please correct English

L692 “reliable data” too generic

REFERENCES

Rubin, D. B.: Inference and missing data, Biometrika, 63, 581–592, 1976.

Little, R. J. A. and Rubin, D. B.: Missing Data in Experiments, in: Statistical Analysis with Missing Data, pp. 24–40, John Wiley & Sons,

Ltd, https://doi.org/10.1002/9781119013563.ch2, 2014.

van Buuren, S.: Flexible Imputation of Missing Data, Second Edition, Chapman and Hall/CRC, Boca Raton, 2 edition edn., 2018.

Bessenbacher, V., Gudmundsson, L. And Seneviratne, S.I: CLIMFILL v0.9: A Framework for Intelligently Gap filling Earth Observations, GMD (in review), https://doi.org/10.5194/gmd-2021-164

Dorigo, W., Wagner, W., Albergel, C., Albrecht, F., Balsamo, G., Brocca, L., Chung, D., Ertl, M., Forkel, M., Gruber, A., Haas, E., Hamer, P. D., Hirschi, M., Ikonen, J., de Jeu, R., Kidd, R., Lahoz, W., Liu, Y. Y., Miralles, D., Mistelbauer, T., Nicolai-Shaw, N., Parinussa, R., Pratola, C., Reimer, C., van der Schalie, R., Seneviratne, S. I., Smolander, T., and Lecomte, P.: ESA CCI Soil Mois- ture for improved Earth system understanding: State-of-the art and future directions, Remote Sensing of Environment, 203, 185–215, https://doi.org/10.1016/j.rse.2017.07.001, 2017.