Response to Comments from Reviewer 1

Note that the reviewer's comments are in italic black, and responses in blue.

We thank the reviewer for his detailed and helpful comments. Here, we carefully address the issues raised by the reviewer to dispel his/her concerns about the added innovation obtained from the study.

1) I am not sure about the utility of using in-situ soil moisture datasets in operational applications. If there are station-based observed datasets, then why not directly use them rather than using other datasets? Afterall, station-based datasets are used in soil moisture validation studies and arguably they represent the best moisture conditions on the ground.

Although station-based datasets are widely used in soil moisture validation, this does not mean station measurements are perfect and have no limitations. However, this is the best available source of information for validation.

In fact, each source of soil moisture information has its advantages and disadvantages. The advantage of *in situ* observations is that they are the only source of ground soil moisture information, and thus are often used as a benchmark for models. The primary disadvantages of *in situ* observations are: 1) sparse spatial density, and 2) limited spatial representativeness. Similarly, model and satellite remote sensing soil moisture have the advantage of representing a larger spatial area and, for the most part, a finer spatial resolution. Of course, the primary disadvantage of models and satellites is that they are indirect representations of soil moisture variations, and not direct measurements. Therefore, these three soil moisture data sources are complementary and, as we show in this study, can be combined to improve soil moisture monitoring.

If the added benefit from blending in-situ datasets is about the regions lacking in-situ observations while dense-networks are present around, then this significantly limits the applicability of the introduced study to regions with dense soil moisture networks (e.g., Oklahoma).

The added benefit from blending in-situ datasets into remote sensing or model simulations is it improves the overall accuracy. We show in this study that Kriging (spatial) interpolation performs best with dense station networks, and therefore presents the greatest advantage in areas of relatively dense *in situ* networks.

On the other hand, there are not many such dense networks that facilitate retrieval of soil moisture by blending information coming from station-based observations and other ancillary datasets (e.g., remote sensing-based precipitation or soil moisture estimates). If this is the case, then 1.a) the motivation of the study & the structure of the introduction section should be given accordingly (i.e., why soil moisture retrieval over Oklahoma-like regions is very important) and 1.b) the study area should be re-selected accordingly by excluding the regions not having ground-stations (i.e., the study area could be limited to state of Oklahoma having ~ 180,000 km2 area + region laying *between 100W-103W & 33N-36N having ~ 90,000km2 area, rather than total 1,150,000km2 area used in this study).*

We agree with the reviewer that "why soil moisture retrieval over Oklahoma-like regions is very important" should be addressed in the introduction section and we will revise it accordingly. However, we respectfully disagree with the suggestion to select a different study area. The current study area includes both dense and sparse networks, which allows us to examine the influence of station density and station representativeness on the blending results (line 705-750). By including areas of dense and sparse *in situ* stations, we can conclude from Fig. 10 and Fig. 12 that when sampling sites are less representative and sparsely distributed, adding an extra source of soil moisture information (e.g., the product of AVE(K,N)) significantly improves the accuracy (line 744-746).

For example, there are regions without any station data over south of Texas; I cannot imagine the sparse networks outside of these regions will add any useful information to these regions lacking any in-situ data.

In this study, we show that blending *in situ* and model soil moisture data provides the most accurate and consistent depiction of soil moisture conditions. Not surprisingly, the benefits are greater when the underlying *in situ* station network is denser; however, even regions with low station density are improved via the blending of the *in situ* and model datasets. Our study (Fig. 10) has shown that in places where in-situ measurements are sparse, such as the central to the south of Texas and east Arkansas, the blending of the *in situ* and model dataset (AVE(K,N)) has similar and sometimes smaller MAE than the *in situ* (K-API) datasets. This indicates when in-situ measurements are sparse, using additional sources of soil moisture information (such that from NLDAS) may help to increase the accuracy (line 707-710).

In addition, based on the First Law of Geography, "everything is related to everything else, but near things are more related than distant things", the sparse network will add the most helpful information to the regions lacking any in-situ data. This is also the foundation of spatial interpolation methods, which is to use the nearby information to estimate the values of unknown locations.

2) Here in this study only simple merging and least squares merging methodologies are compared. REV and TCA are error variance estimation-related methodologies (i.e., not blending), while kriging is more like a spatial interpolation rather than being blending methodology that facilitates merging of two, three, four datasets, unlike least squares or simple merging methodologies. Given earlier studies have already performed least squares - simple merging methodology comparison (e.g., as cited by authors, Yilmaz et al., 2012 have already done it), I don't agree with the statement that "comprehensive evaluation of different data blending methods" has not been implemented. Here, if the contribution is about "investigation of the impact of estimated error variance information over the blending", then perhaps this should be reflected in the title as well as the motivation given in the introduction.

Thank you for the helpful suggestion. We will change innovation point (2) to "investigation of the impact of estimated error variance information over the blending" and add a description of this motivation in the introduction.

3) 3.a) It seems apples and oranges are being compared. Even though native datasets and anomaly information are utilized separately in the merging methodology, their comparisons on an equal ground have not been performed. It does not make much sense to compare the errors of the native datasets and errors of the anomaly components of the same datasets (i.e., native time series are composed of two independent components, called in this study anomaly and climatology). If the goal is to obtain a product that is more like raw-product in nature, then the merged anomaly product should be converted back to the space of native product by adding the already-subtracted climatology component, and then the errors of this anomaly + climatology merged product could be directly compared against the errors of the native product-like merged estimate.

The direct comparison between the three data formats (absolute values, anomalies, percentiles) is to determine the impact of measurement units on soil moisture blending, which is also our third innovation point.

In this study, the MAE, RMSE, and NSE were used to compare different formats on an equal ground. Line 649 to 664 discussed the comparison between the native and anomaly data and examined the impact of data format on hybrid results. The MSE_MD^2 and MSE_VAR are also used to evaluate the errors due to the mean/bias and errors due to the variance, respectively. It is found that the bias (MSE_MD2) in the anomalies and percentiles are close to zero, but not in the raw data. This indicates both anomalies and percentiles are useful for removing the systematic bias between different datasets. However, when using absolute soil moisture data (VWC) the bias-related errors are not removed from the final datasets.

We also agree that "Comparing the errors of anomaly + climatology merged product against the errors of the native product-like merged estimate" is an alternative way for data assessment. However, the estimation of climatology may introduce another source of error or uncertainty. To serve our purpose, that is to provide guidance and references on the data format for data sharing and distribution, the three formats were compared in our study, and differences among the measurement units were investigated using different indicators (MAE, RMSE, NSE MSE_MD2 and MSE_VAR).

3.b) I am not convinced that merging native products without a proper rescaling methodology is justifiable. For the last two decades numerous of studies have clearly shown that there are systematic differences between the statistics (e.g., mean and standard deviation) of soil moisture estimates (e.g., Dirmeyer et al., 2004; and Reichle & Koster, 2004). These systematic differences should be removed via certain rescaling methodologies before they could be merged (Afshar et al., 2019). The results shown in this manuscript are also very consistent with these earlier studies: absolute value merging (i.e., no rescaling before merging) yields the worst performance (nash value 0.25) compared with rescaled versions via anomalies and percentiles (nash values 0.60 or *higher). Accordingly, all "absolute value" related investigations are redundant, hence should be removed from the study.*

The results of previous studies, included those cited by the reviewer, inform our decision to include an absolute value investigation. We demonstrate that rescaling prior to dataset blending results in significant improvement over the blending of absolute values. This agrees with previous studies. This particular analysis is not redundant, but is instead complimentary of previous studies, and is just one of many findings in this study. Therefore, we believe that our study is stronger and more impactful with the inclusion of the absolute value-related investigations.

4) The manuscript states "A simple and operational methodology is still needed for accurate daily soil moisture mapping with high spatial resolution". I strongly believe "simple methodology for accurate daily soil moisture mapping with high spatial resolution for operational applications" exists. Authors should clearly state what is missing in the established literature with more detail.

Thank you for your suggestion. Indeed, this statement was too general. We will revise and instead argue that a simple and operational methodology is still needed to blend multiple, diverse sources of soil moisture data in order to produce the most accurate, high resolution soil moisture maps possible.

5) There are many unjustified/redundant statements in the manuscript all over. They hinder the impact of the delivered messages. - Lines 66-79, following earlier studies (Dirmeyer et al., (2004) and Reichle & Koster, 2004), there is no use in stating the facts that there are systematic differences between the model-based soil moisture products.

Thank you for your suggestion. The statement on the systematic differences between the model-based soil moisture products (Line 66-79) will be condensed in the revised manuscript.

- Lines 121-123, "However, none of them, at least by themselves, are adequate for providing accurate soil moisture data at high temporal and spatial resolutions.". Noah model runs at 1km spatial resolution and 1-hour temporal resolutions exist (e.g., LIS model-based runs can simulate soil moisture at 1km spatial resolution and 1-hour temporal resolution globally). Is it not adequate in terms of spatial and temporal resolutions?

We agree that this statement was not specific enough. Noah 1 km runs are adequate in terms of spatial and temporal resolution; however, the output is entirely model simulated and therefore does not contain any ground truth. The inclusion of *in situ* data with high resolution model output, therefore represents a significant advancement in soil moisture monitoring. We will clarify this statement in the revised text.

- Lines 123-125, "Therefore, it is useful to combine these three independent data sources to capitalize on the strengths of each and to generate an optimal soil moisture product to facilitate

real-world applications.". Without merging datasets realworld applications can not be facilitated? So, does it mean model runs with 1km spatial resolution and 1-hour temporal resolution is not sufficient?

Our statement here is not to downplay the importance of other products, but to stress the improvement that can be achieved by incorporating in-situ measurements. This is stated clearly, and nowhere in the manuscript do we imply current operational products are unfit for real-world applications.

- Lines 137-138, "Current studies mainly focus on combining modeled and RS soil moisture, rather than combining all three sources (modeled, RS and in-situ)", over which locations? Given such in-situ datasets are limited what is the added benefit obtained from in-situ observations globally? Should we only focus on local studies? But, if we focus on local studies and know the in-situ data, then why merge such very high-accuracy products with datasets with much lower accuracy?

Thank you for the series of questions. The locations of the current studies will be investigated and added to the revised manuscript.

The benefit of adding *in situ* data to current products is improved accuracy. The reason that *in situ* data has not been widely used or added is due to the sparse soil moisture network. Once the soil moisture network has developed to the same scale and density as the weather stations, the impact and importance of soil moisture stations will be huge both locally and globally.

Our final goal is to obtain improved accuracy of soil moisture produce at global scale. To achieve the big goal, the experiment is first conducted at local scale, and future research is required to apply the methodology at global scale.

- Lines 515-516, "Our study also demonstrates that the measurement units (Fig. S4) do not impact the relative relationship (error ranking) between the different datasets". I don't really see how this result is obtained from this figure.

Thank you for the comment. We will add details and revise the sentence to "Our study also demonstrates that the measurement units do not impact the relative relationship (error ranking) between the different datasets (Fig. S4). Take the first column of Fig. S4 for example, no matter whether anomalies (top subplot), percentiles (middle subplot), or Kriged SM (bottom subplot) is used, the In-situ presents the highest error, followed by SMAP, and NLDAS has the lowest error among the three. The same rule can be found when examining the second to fourth columns in Fig. S4."

- Line 685, "...sites that are less temporally representative sites ...". "Temporally representative" phrase should be elaborated.

Thank you for the comment. The "temporal representativeness" has been clarified in Section 2.5.3, therefore we will modify the sentence to "As mentioned in Section 2.5.3, a temporally representative site is one with a small mean bias or a low value of ITS".

6) "In this study, a set of k values (from 0.80 to 0.99) is tested to determine the optimal k value that results in the highest correlation between API and soil moisture based on 215 stations." But this is clearly overfitting: k values are fit to yield API values that give highest correlations, while these API products are later used to obtain gridded products which are later validated using the same in-situ datasets that the API is calibrated against.

Thank you for this comment. In this study, the 215 stations were used as a whole to give an **overall** estimation of the **k** over the study area. Later, the gridded product from **API** is validated on 127 (60%) stations. The modeled parameter (k) and the validated parameter (API) are different. Therefore, we think the overlapping issue may be negligible. Later, we will confirm this idea by conducting a sensitivity analysis of k values with different station numbers.

7) Equation 14 introduces a rescaling step before the triple collocation given in equations 15-17. On the other hand, Stoffelen (1998) clearly introduced another methodology to rescale the datasets before the error estimation (equation 2 in the study of Stoffelen, 1998). Without proper rescaling step, the triple collocation methodology (i.e., equations 15-20) would be void. Accordingly, the use of equation 14 before equations 15-20 is absolutely not acceptable (i.e., if the original rescaling steps given by Stoffelen, 1998, is used, then equation 14 is redundant). The original methodology introduced by Stoffelen (1998) must be followed.

Thank you for your suggestion. The rescaling methods used by this study has been successfully applied with TC by Dorigo et al. (2010). In addition, the new research from Afshar et al. (2019) also revealed that "Variations in rescaling methods have only a small impact on the precision of pair fused products". Therefore, we expect little improvement from using other rescaling methods.

Dorigo, et al. (2010). Error characterization of global active and passive microwave soil moisture data sets. Hydrology and Earth System Sciences, 14, 2605--2616

Afshar et al. (2019). Impact of rescaling approaches in simple fusion of soil moisture products. Water Resources Research, 55, 7804–7825.

8) "Our preliminary results showed that the choice of reference dataset did not impact the final results, thus the RK-gridded soil moisture is selected as the reference dataset in this study." There is plenty of literature available showing that reference dataset selection matters in rescaling soil moisture products (Afshar et al., 2019). Accordingly the manuscript should show which "final results" are not impacted.

Thank you for the great suggestion. We will cite the work from Afshar et al (2019) and revise the statement to "Our preliminary results showed that the choice of the reference data set does not influence the relative magnitude of the errors. In addition, Afshar et al (2019) revealed that a more precise reference product yielded a higher precision merged soil moisture product. Therefore, the in-situ gridded soil moisture is selected as the reference dataset in this study.”

9) In order to show the RK-gridding impact, manuscript also show the performance of merging SMAP, NLDAS, API products (i.e., API itself without blending with in-situ data using kriging). Only after this step the manuscript can claim some added benefit obtained from kriging methodology (i.e., otherwise the added benefit seems to stem from API dataset).

Thank you for this good suggestion! The examination (comparison between API and blended API) will be added in the revised manuscript.

10) It seems “displacement of autocorrelation” lies at the hearth of the “Relative error variance” methodology, while equations 21-23 does not really show how exactly it is calculated. Frankly I did not understand “The displacement of the autocorrelation ngamma(ntau) at ntau = 0”. I am puzzled with “autocorrelation with 0-lag” (i.e., to me autocorrelation at 0-lag should be equal to 1; this clearly shows I am missing something while the manuscript does not help me). Step by step instructions are needed (i.e., given I have soil moisture time series, how exactly I should code this? The details should give this much information).

Thank you for the suggestion, we will further clarify the computation of Relative Error Variance in the revised text. Briefly here, a soil moisture time series with no variance attributable to measurement error will, in theory, have a 0-lag autocorrelation of 1. However, when fitting a linear regression against the autocorrelation from lag 1 to x, the estimated 0-lag autocorrelation will deviate from 0 in proportion to the relative error variance. This is the theory behind the Dirmeyer *et al.* (2016) methodology.

11) Giving too much emphasis to the differences between the raw products (Lines 465-480) does not make much sense to me given earlier studies already says it (Dirmeyer et al., 2004, and Reichle & Koster, 2004). I think this paragraph is redundant.

Thank you for your good suggestion. We will replace this paragraph by citing Dirmeyer et al (2004)’s work.

12) Fig 4 shows four different TCA results using different products (assuming different reference datasets are selected in each of these TCA calculations). Accordingly, comparison of absolute value TCA errors retrieved using different reference datasets does not make much sense. I strongly recommend authors to have a look at the study of Draper et al., (2013) in this context.

Thank you for the reviewer’s suggestion. Although Draper et al., (2013)’s paper is a good one, we have different focus. The purpose of Draper et al. (2013) paper is to estimate the errors (RMSE) of two remote sensing soil moisture products (ASCAT and AMSR-E) based on TC. They were dealing with TC using fixed (only 3) parent triplets (ASCAT, AMSR-E and in-situ). However, in our study, Fig. 4 (a)-(d) aims to examine the performance of TC when the parent triplets are changing. Here, we have four parent products (in-situ, K-API, SMAP, and NLDAS), and each time we selected three from them for TC calculation. Fig. 4 (a)-(c) showed the TC estimation based on the same (in-situ) reference dataset, while Fig. 4 (d) used the K-API as reference data.

Our results indicated the error estimated from TC depends on the parent triplets used. Changing the parent triplets changes the magnitude and error ranking of the parent datasets, even if the same reference dataset was used (such as Fig. 4(a)-(c)). Therefore, the relative errors estimated from TC are sensitive to the choice of input datasets, and caution should be taken when selecting the input datasets for TC analysis. This provides a new angle to examine the TC method and provide guidance for its future application.

13) The word "significance" has been written > 25 times, while not a single sentence is written how exactly these "significance tests" are applied (i.e., at what confidence level, which significance test?).

Thank you for the suggestion. The confidence level of the significant test will be added throughout the text in the revised manuscript.

15) The abbreviations "RK-gridded" and "K-API" are used interchangeably. Consistent use of abbreviations is needed.

Thank you for the suggestion. We will check our abbreviations and make sure they are consistent throughout the text in the revised manuscript.

16) "The in-situ measurements cannot be considered the "truth" because they are point measurements that may not reflect the soil moisture value for each 4 km grid cell." But almost all soil moisture validation efforts (e.g., Jackson et al., 2010, https://doi.org/10.1109/TGRS.2010.2051035) and this present manuscript are done using in-situ based observations. I don't agree with this statement.

In situ soil moisture observations are arguably closer to actual soil moisture conditions within a meter or two of the sensor than model or satellite estimates. The sensors used at these operational stations have inherent measurement errors and failure rates, and therefore cannot be considered an absolute truth. The only "truth" is a gravimetric sample.