RESPONSE TO REVIEWER #1

Author's response: We really appreciate the comments from the reviewers and note some common themes: (1) *why were the parameters and forcing choices made*? R1.1, R1.2, R1.4, R2.6, R2.7 (2) *why were the evaluation methods chosen* R2.5, R2.8, R2.9. (3) some general questions that we think could be clarified by the motivation, and our interpretation of the scope of a data descriptor (4) *Addition of a clear 'limitations and future work' section* (5) Editorial comments: Figure legibility, figure placement, edits to intro and abstract.

Dear Reviewer #1, Thanks for the positive comments and we address your remarks and suggested revisions in the following.

1. It would be useful for the modeling community to understand a bit more about the data choices summarized in Table 1. Why were each of these parameter datasets selected? For example, the FAO soil texture dataset has been replaced by ISRIC in many applications, and the NCEP vegetation fraction dataset is a low resolution climatology. I don't ask that the authors change these settings, but given the status of FLDAS as an operational LDAS system with global scope I expect that the authors are in a good position to provide readers with some guidance regarding the choice of parameter sets. (Also: a minor note on Table 1—the row labels could be improved: why is the FAO soil data simply listed as "parameters" and the snow albedo simply listed as "albedo"? It looks like words were dropped.)

Thanks for this comment. We propose to add some variation of the following text:

"The parameters and specifications listed in Table 1 are largely default settings defined by the Noah LSM community (NCAR Research Applications Library, 2021). Ongoing research aims to identify where model output performance can be improved with parameter updates. Evaluating parameter updates had similar challenges as evaluating input forcing described in Section X: without reliable reference data it is difficult to determine a "best" input. For example, we have explored changing soil parameters from FAO to International Soil Reference and Information Centre (ISRIC) SoilGrids database (Hengle et al. 2017). However this change did not result in improvements in streamflow statistics in Southern Africa, nor for soil moisture anomaly's ability to represent drought events. We expect similar results in Afghanistan where e.g., streamflow will be sensitive to a change in soil parameters and the challenge will be to obtain referenced data to evaluate if there is an improvement. Moreover, our model runs at 0.1 and 0.001 degrees may not fully exploit the added value of the 250m soil grids (noted in Ellenburg et al. (2021)) for a LIS application in East Africa.

Vegetation parameters are also potential sources of improvement whose importance to LDAS hydrologic estimates has been highlighted (e.g., Miller et al. 2006). We have found the NCEP estimates of green vegetation fraction (GVF) to be sufficient for this configuration of Noah 3.6, due to lack of demonstrated improvements with e.g. a time series of green vegetation fraction derived from NDVI which did not improve representation of droughts. However, Future FLDAS

Global and Central Asia, versions will be run with Noah-MP (Niu et al. 2007) which has vegetation options beyond that of Noah 3.6, and relies on Leaf Area Index rather than GVF, which will open up possibilities for choice of datasets to meet our application needs and potentially improve representation of the water balance.

Re: Also: a minor note on Table 1… Thank you, we have updated the rows in Table 1. Changing "Parameters" to "Soil Parameters", and combined average and max albelo into "Albedo"

> 2. Section 2.3: The authors explain how they used the different precipitation datasets, but I'm confused about the rationale. IMERG is not a particularly new product—it's been available for several years. Why does the system still use GDAS precipitation and only include IMERG as a data comparison? The IMERG Late (or Early) runs would have low enough latency for real-time monitoring applications, and the authors later note that they plan to integrate IMERG into the system. Is there a reason why this isn't already done? For example, some practical or product continuity advantage to using GDAS instead of IMERG?

Thanks for this question, and fits with themes we've identified in the reviews (1) why were the inputs chosen (2) better description of limitations and future work. We also propose to add text to the following sections:

Section 1.2 Meteorological Background
… our choices of inputs must meet the following criteria: (a) provide a long historic record for contextualizing estimates in terms of departures from the mean (anomalies) (b) are low latency (< 1-month) for timely decision support (c) are familiar to the FEWS NET user-community and (d) prior evaluation by our team and the broader community.
In addition to these products (GDAS, CHIRPS, MERRA-2) the Integrated Multi-satellite Retrievals for the Global Precipitation Mission (IMERG), a NASA precipitation (Huffman et al., 2020) has emerged as a precipitation product that meets these requirements, since its period of record was extended back to 2000 as of version 6 which was released in 2019.

Section 3.4 Limitations and Future work:
As of 2019, with the release of IMERG v6, these data go back to 2000 as well. Prior to this change we began comparisons within our framework, described in (Kirschbaum et al. 2018) and found encouraging results. At that time however, the period of record was a limitation for computing anomalies. We now have an adequate period of record, and IMERG will be part of the upcoming FLDAS-Global and FLDAS-CA releases. We are also encouraged by the quality of IMERG: For example, Sarmiento et al. (2021) further explored the qualities of IMERG compared to CHIRPS over CONUS.

Other projects are also informing our understanding of how IMERG will perform within our system. One recent attempt to improve meteorological inputs in the region is from Ma et al. (2020) with the development of the AIMERG dataset that combines IMERG Final with APHRODITE (Asian Precipitation - Highly-Resolved Observational Data Integration Toward

Evaluation) rain-gauge derived product (Yatagai et al., 2012). Another promising development is CHIMES (Funk et al. 2022) derived from both CHIRPS and IMERG, whose developers have been exploring the strengths and limitations of these two datasets and how to fuse them to produce an optimal product.

So, great point, we're certainly moving in that direction to incorporate the benefits of IMERG into the products co-authors are developing for the food/water security user community. We anticipate you may see these products on the USGS website in the next year, as well as associated technical documentation.

3. Figure 6: Can the authors comment on the fact that the coarse resolution global run appears to do better than the high resolution CA run in this data comparison? I find it surprising, given the presumed topographic sensitivity of SCA.

Thanks for this comment. The categorical statistics do indicate that Central Asia (GDAS) tends to have both a higher probability of detection AND false alarm rate, indicating higher mean than MODIS SCF and Global (CHIRPS). Figure 6 emphases the high mean and apparent better correspondence of the global run. This may be further exacerbated by the spatial averaging over the basin.

For an alternative perspective we conducted a pixel-wise NIC comparison of Central Asia, Global run's estimates of evapotranspiration vs SSEBop ET. We found that in general, GDAS derived estimates of ET consistently performed better over Afghanistan in terms of pixel-wise correlation with SSEBop ET (Figures below). The modeled estimates of surface soil moisture vs SMAP were more mixed depending on the month.

In general, given the lack of clarity on "best" product, our approach is to highlight the potential benefit of CHIRPS as highlighted in other literature (as well as its familiarity to the food security community), while GDAS has a benefit for our applications of being a low latency and higher spatial resolution for routine monitoring.

Given the potential for confusion, rather than adding clarify/information to our description we will *remove Figure 6.*

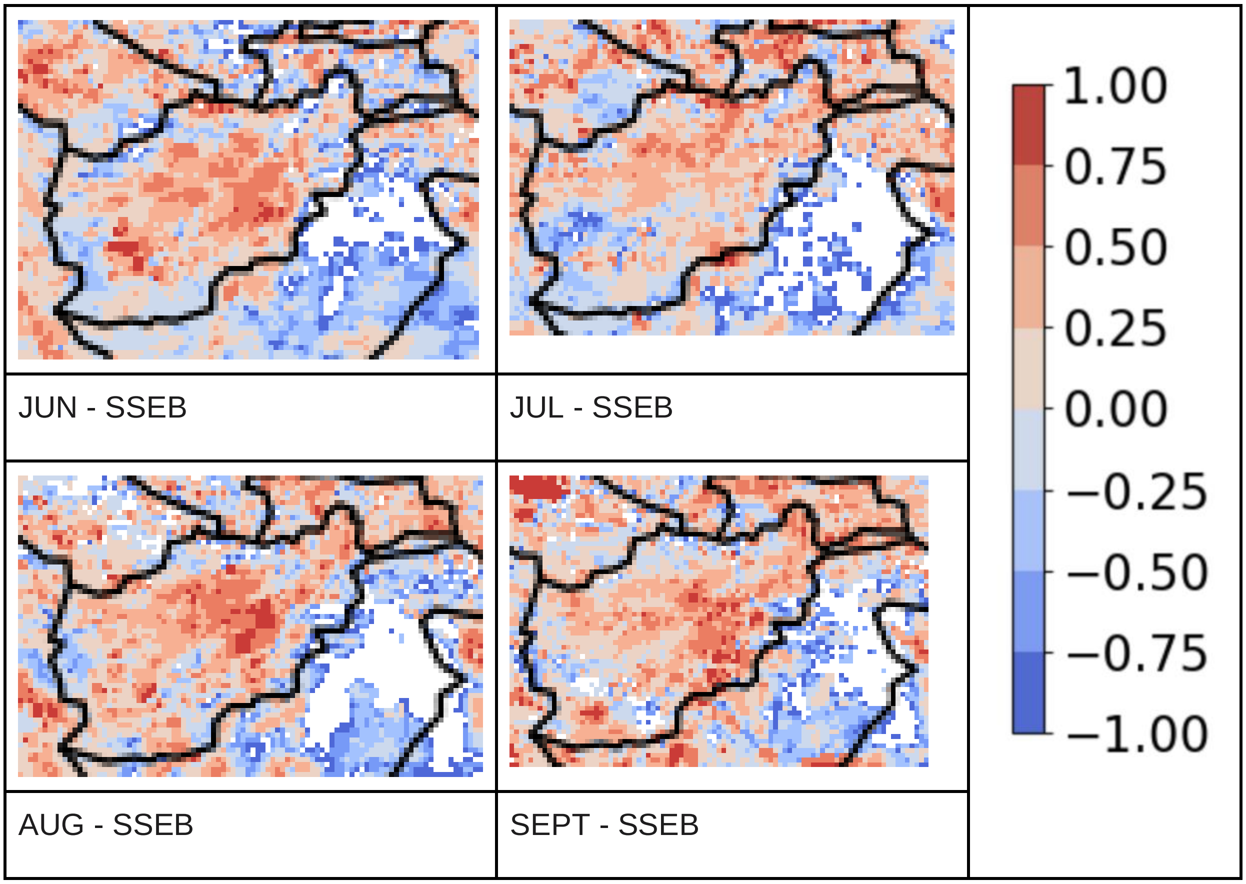

Figure R1. Normalized Information Contribution (NIC) comparison between Global (CHIRPS), and Central Asia (GDAS) monthly evapotranspiration estimates and SSEBop. Red indicates the reference dataset (SSEB) has relatively higher correlation with the Central Asia datastream and blue indicates relatively higher correlation of the Global runs.

Normalized Information Contribution (NIC) metric (following Sarmiento et al 2021) as used to compare the differences between the simulations using the Simplified Surface Energy Balance for operations (SSEBop) (Senay et al. 2013). The NIC used the simulated monthly anomalies for every month and correlated those with the SSEB monthly anomalies.

1. The authors note various limitations and potential areas of improvement throughout the paper. I would find it useful for this information to be included in a short section near the end of the paper on "Limitations and Future Work" that could describe ongoing FLDAS development activities. While the future work isn't a necessary component of this data description paper, it would be valuable for the reader to have this information when considering adopting FLDAS to support research or operations.

Good suggestion. We have changed Section 3.4 Limitations and Future Developments to include more discussion beyond Table 3 'pros and cons'

We now include discussion of the IMERG dataset, and future plans for continued evaluation and incorporation into future FLDAS versions (see response #2)
As previously mentioned regarding parameters, FLDAS Central Asia and Global will be transitioning to Noah-MP. This will allow for improved representation of glaciers and groundwater. This will also necessitate the use of different parameters e.g., LAI, as well as the potential to explore different parameter sets like ISRIC soils.

We note that our particular configuration has been developed within a specific set of constraints and for a specific user community that gives weight to length of record, latency, and familiarity with or compatibility with other routinely used products. We encourage the data user to consider how these factors may influence their particular use of and interpretation of the dataset. We also welcome feedback regarding the performance of these data when confronted with other questions, regions and reference datasets.