We would like to thank the reviewers for the thoughtful comments and suggestions. See below for our responses to all comments. The reviewer' original comments are in read italic fonts and our responses are in regular black fonts.

Anonymous Referee #1

This is a well-written document describing the LVT, a comprehensive tool for evaluation of land-hydrology models which includes a number of insightful methods of model evaluation, from model parameter estimation and evaluations against satellite data to the characterization of uncertainty diagnostics (among the many topics they describe with interesting, useful and informative examples), using various types of data sets (single point as well as remote sensing/spatially-distributed) in the LVT analyses. Combined with LIS, this is truly an end-to-end system for land modeling and evaluation.
Only a couple of (minor) questions/comments:

section 5.1 An end-to-end example of the MDF paradigm "Figure 3 shows a comparison of the mean diurnal cycles of latent and sensible heat fluxes... The calibration of model parameters helps in improving the model performance..." » INTERESTING –WHAT PARAMETERS? (JUST CURIOUS) THIS IS NOT REALLY SO IMPORTANT SINCE PRESUMABLY, USING LVT THE QUESTION OF ROBUSTNESS OF THE CALIBRATED PARAMETERS FOR OTHER SEASONS/REGIONS MAY BE ASSESSED.

In this example, we calibrated 29 Noah model parameters, including vegetation and soil properties. Section 5.1 has been modified to mention this additional information.
The reviewer is correct that in this example, we do not make any attempt to verify the robustness or the transferability of the estimated parameters. The experiment is used to illustrate that LVT can be used to evaluate the impact of model changes in a systematic manner, as envisioned in the MDF paradigm.

section 5.2 Example of model evaluation against satellite data "High values of POD and low values of FAR are observed over the Central Highlands region of the domain, suggesting a high degree of accuracy of model snow cover estimates over these areas. Over the northeast parts of the domain ... model simulations are less accurate..." » AGAIN, PRESUMABLY SOME LVT OPTIONS MAY BE USED TO INFER THE POSSIBLE EXPLANATION(S) OF THE PERFORMANCE NOTED HERE (SNOW ACCURACY DIFFERS BY REGION IN THE DOMAIN). (AGAIN, JUST CURIOUS)

We agree with the comment. In fact, a separate manuscript that examines the terrain effects on snow simulations has been submitted to the Journal of Hydrometeorology. In this paper, we employed the features of LVT to examine and stratify the performance of the model as a function of terrain characteristics and snow season. A more detailed analysis is omitted here since the focus here is simply to demonstrate the use of satellite datasets for model evaluation through LVT.

Anonymous Referee #2

This paper describes a new model evaluation system, called the Land surface Verification Toolkit (LVT), designed to operate in conjunction with NASA's Land Information System (LIS). LVT uses a large collection of in situ, remote sensing, reanalysis, and data assimilation-based model results to judge the fidelity of hydrological model results from multiple land surface models operating within LIS or results translated into LIS-style outputs. LVT offers a variety of metrics, including those for point comparisons, spatial correspondence, ensemble and uncertainty measures, information theoretic metrics, and spatial scale impacts on model performance.

§1 is well-written and appropriately motivates discussion of LVT.

In §2, the authors describe several community-wide efforts—citing GSWP, ALMIP, and C-LAMP—as providing "benchmarks for the simulation of continental scale water andenergy budgets." However, the authors fail to mention that C-LAMP also included bio- geochemical model evaluations of global forest phenology, global primary productivity, CO2 seasonality, and regional carbon stocks and dynamics. Overall, §2 offers a good background of previous and current model evaluation efforts and provides the context for development of LVT.

We thank the reviewer for the comment. Section 2 has been modified to include these additional details regarding C-LAMP.

While the system is conceptually well described in the paper, architectural details are lacking, resource requirements are not described, and computational performance is not discussed. In particular, §3 describes the three-layer structure of LVT, but offers no details on how the Analysis Metrics and Observations in the Abstractions layer are implemented in the Fortran 90 and C languages. Are Metrics templated for ease of extensibility? ESMF would seem to be a heavy-weight solution to provide only clock/time, configuration, and logging infrastructure in the Core Structure and Features layer. Does it also provide the geospatial transformations functionality, or are other packages employed for that purpose? What are the typical memory and CPU requirements of the LVT package, and what are its computational and I/O performance characteristics? Does LVT employ shared- or distributed-memory parallelism?

We have modified section 3 to include the additional details on the design. They are again summarized below:

The virtual function tables written in C enable the Abstractions layer.

The metrics are "templated" for extensibility. All metric implementations extend the "abstract" definition of a metric (defined in the Abstractions layer), as noted in the first paragraph.

Currently, the ESMF usage in LVT is limited to the infrastructure layer of ESMF. We also expect to use the geospatial transformations and communications layer of ESMF in the newer versions of the ESMF library.

In the examples presented in the article, LVT has been used as a serial code. The parallelism support is currently in development. (The summary section has been edited to

mention this detail).

The memory and CPU requirements and the corresponding performance of LVT are largely determined by the analysis domain, the datasets and the metrics being computed. These requirements increase with large geospatial domains with high resolutions. As a result, it is difficult to describe the "typical" requirements. We have modified section 4.3 to mention these details.

In §4, the authors describe a philosophy of preserving datasets in their native formats, but it is not clear if LVT reprocesses those datasets every time an evaluation is executed to make them consistent, spatially and temporally, with model results. For high resolution remote sensing datasets, generating products for spatial comparison with low resolution model grids and low frequency output could take considerable compute time. How are these issues handled? §4.2 offers useful motivation for careful consideration of appropriate analysis metrics.

By default, LVT reads, interpolates, reprojects and subsets the datasets from their native formats to the analysis domain. This is handled in memory and is done every time LVT is run. The reviewer is correct that the combination of coarse resolution analysis domains and high resolution datasets could be costly in terms of the compute time. To circumvent this limitation, LVT provides a "data processing" run mode where it would simply perform the data transformation operations and write the processed files to disk to be read by a subsequent analysis run. We have modified section 4 to clarify these additional features.

The examples provided in §5 are extremely valuable, but no details are provided about how graphical diagnostics are produced. Are one or more graphical packages employed by LVT, or is the user left to generate graphics from numerical output from LVT?

LVT does not include any graphical packages in it. LVT writes the analysis outputs in the user desired formats (ASCII, GRIB, Binary, NETCDF) and the generation of graphics is left to the user. We have modified section 4.3 to explicitly mention this detail.

For the MDF paradigm example described in §5.1, it is slightly misleading that the authors suggest "calibration of model parameters helps in improving the model performance, by correcting both these systematic biases", when the sensible and latent heat biases are really the result of a single bias in energy partitioning.

We agree and have modified the statement to say " …..by correcting the systematic bias in energy partitioning"

§6 does a good job of summarizing the key points of the paper.
This paper does not describe the relative importance of different metrics or if LVT provides a mechanism for weighting metrics or scoring models based on a collection of these metrics, incorporating model and observational uncertainties. Does LVT offer such a

mechanism for judging the overall performance of models for one or more specific application areas?

The reviewer is correct that no attempt is made to compare the relative performance of different metrics and currently no mechanisms are available for weighting metrics or scoring models. In fact, these goals will be pursued in the continued development of LVT. We have added these details in Section 6.

A few potential typographical errors were noted. On page 244, line 20, the article "the" may be extraneous before "near-optimal performance." On page 246, line 13, "of" is missing before "such sensitivities." On page 248, lines 13 and 14, the "in" and "to" that span those lines might more appropriately be one word. On page 248, line 28, an article ("a" or "the") may improve the readability before "sum of orthogonal components". On page 240, line 11, "representation" should probably be plural.

Thank you for these corrections. The manuscript has been updated with these changes.