Evaluation of High Mountain Asia - Land Data Assimilation System (version 1) from 2003 to 2016, Part II: The Impact of Assimilating Satellite-Based Snow Cover and Freeze/Thaw Observations into a Land Surface Model

Yuan Xue[1], Paul R. Houser[1], Viviana Maggioni[2], Yiwen Mei[3], Sujay V. Kumar[4], and Yeosang Yoon[4,5]

[1]Department of Geography and GeoInformation Science, George Mason University, Fairfax, Virginia, US, 22030.

[2]Sid and Reva Dewberry Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, Fairfax, Virginia, US, 22030.

[3]Sid and Reva Dewberry Department of Civil, Environmental, and Infrastructure Engineering, George Mason University, Fairfax, Virginia, US, 22030. Now at School for Environment and Sustainability, University of Michigan.

[4]Hydrological Sciences Laboratory, NASA/GSFC, Greenbelt, Maryland, US, 20771.

[5]Science Applications International Corporation, McLean, Virginia, US, 22102.

Key Points:

- The skills of the High Mountain Asia - Land Data Assimilation System (version 1) with and without multi-variate assimilation are presented.
- The assimilation enabled experiments generally perform better than the no-assimilation counterparts.
- The assimilation enabled experiments mitigate the negative effects arising from the fixed long-term precipitation correction factors.

Corresponding author: Yuan Xue, yxue4@gmu.edu

Abstract

This second paper of the two-part series focuses on demonstrating the impact of assimilating satellite-based snow cover and freeze/thaw observations into the hyper-resolution, offline terrestrial modeling system used for the High Mountain Asia (HMA) region from 2003 to 2016. To this end, this study systematically evaluates a total of six sets of 0.01° ($\sim$ 1-km) model simulations forced by different precipitation forcings, with and without the dual assimilation scheme enabled, at point-scale, basin-scale, and domain-scale. The key variables of interest include surface net shortwave radiation, surface net longwave radiation, skin temperature, near-surface soil temperature, snow depth, snow water equivalent (SWE), and total runoff. First, the point-scale assessment is mainly conducted via evaluating against ground-based measurements. In general, the assimilation enabled estimates are better than no-assimilation counterparts. Second, the basin-scale runoff assessment demonstrates that across three snow-dominated basins, the assimilation enabled experiment yields systematic improvements in all goodness-of-fit statistics through mitigating the negative effects brought by the fixed long-term precipitation correction factors. For example, when forced by the bias-corrected precipitation, the assimilation-enabled experiment improves the bias by 69%, the root-mean-squared error by 30%, and the unbiased root-mean-squared error by 18% (relative to the no-assimilation counterpart). Finally, the domain-scale assessment is conducted via evaluating against satellite-based SWE and skin temperature products. Both set of domain-scale analysis further corroborate the findings in the point-scale evaluations. Overall, this study suggests the benefits of the proposed multi-variate assimilation system in improving the cryospheric-hydrological process within a land surface model for use in HMA.

1 Introduction

The first part of the study presented the importance of using a hyper-resolution modeling configuration to characterize the cryospheric-hydrological process across the complex High Mountain Asia (HMA) region. It is acknowledged that, although we improve the spatial resolution of the model input and output, model estimates are inevitably imperfect mainly due to limitations such as imperfect model parameterizations and atmospheric boundary conditions (Nandakumar & Mein, 1997; Mendoza et al., 2015; Zheng et al., 2017). Among the surface meteorological data used to drive the hyper-resolution modeling unit, precipitation is the most important mass input variable (Guo et al., 2006;

Yoon et al., 2019). However, all precipitation estimates contain errors and uncertainty, especially in complex terrain (Yilmaz et al., 2005; Maggioni et al., 2017). Land surface modelers seek to use a precipitation estimate closest to the "truth" in their own study domains. However, there is often no clear answer to this question (Kidd & Huffman, 2011; Gehne et al., 2016; Xue et al., 2021).

To overcome the aforementioned model deficiencies, data assimilation (DA) is often used – which is referred to as an approach to constrain physical land surface model derived estimates through the input of the observation. It is assumed that the dynamics that are responsible for a particular process are inherent in the observations (Hofmann & Friedrichs, 2001). By constraining the model with various observations, model estimates can be improved. Xue et al. (2019) successfully demonstrates the efficacy of assimilating the satellite-based freeze/thaw as well as the snow cover product independently into the hyper-resolution land surface model to improve model estimates across HMA for a sample water year via simplistic rule-based direct insertion algorithms. Based on the encouraging results seen in Xue et al. (2019), we want to demonstrate the effect of dual assimilation (i.e., joint assimilation of snow cover and freeze/thaw) in this study. To our best knowledge, there exists no published study performing rule-based dual assimilation for the entire HMA for a relatively long period (e.g., more than 10 years) at a fine spatial resolution (e.g., finer than 5-km).

Using Xue et al. (2021) as a benchmark, in this study, we attempt to address the following science questions: 1) To what extent does assimilation of satellite-based products improve or worsen land surface modeling, compared to ground-based observations or satellite-derived reference products? 2) Can the dual assimilation scheme be beneficial at mitigating systematic biases possibly caused by overly-corrected precipitation? To this end, this study systematically evaluates six sets of 0.01° ($\sim$ 1-km) model simulations at point-scale, basin-scale, and domain-scale. The key variables of interest include surface net shortwave radiation, surface net longwave radiation, skin temperature, near-surface soil temperature, snow depth, snow water equivalent (SWE), and total runoff. The ultimate goal of this research is to evaluate the newly-developed, hyper-resolution High Mountain Asia - Land Data Assimilation System (HMA-LDAS; version 1) from 2003 to 2016. As mentioned in Xue et al. (2021), HMA-LDAS is intended to provide spatially and temporally continuous land surface estimates which are essential for capturing the spatio-temporal evolution of hydrometeorological conditions and their associated pro-

cesses across HMA characterized by complex terrain. Part II, presented in this manuscript, focuses on demonstrating the impact of simultaneously assimilating satellite-based snow cover and freeze/thaw observations into a hyper-resolution (at $\sim$ 1-km spatial resolution) terrestrial modeling system forced by different precipitation forcings.

2 Data and Methods

2.1 Snow cover and freeze/thaw assimilation

In DA enabled experiments, we assimilate satellite-based snow cover and freeze/thaw observations into the model simultaneously. The forward model used here is the Noah-Multiparameterization Land Surface Model (Noah-MP; version 3.6: i.e., without the glacier modeling routine). The satellite-based snow cover products are obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) Snow Cover Daily L3 Global 500-m Grid (MOD10A1, version 6; Hall and Riggs (2016)). Following Rodell and Houser (2004), Arsenault, Houser, De Lannoy, and Dirmeyer (2013), and Xue et al. (2019), direct updates in SWE and snow depth take place daily at 00:00 (UTC) through the DA update analysis step. The rule-based (a.k.a., direct insertion based) snow cover assimilation scheme assimilates daily, binary 0.01° (i.e., same model grid) snow cover maps post-processed from the MOD10A1 product, with ones (i.e., representing snow covered conditions) and zeros (i.e., representing snow-free conditions) for land pixels. Figure 1 shows the spatial map of the total percentage of days with valid MODIS snow cover observations from 01 February 2003 to 30 November 2016 as well as its variability across different seasons for the entire period. Overall, the spatially averaged mean is 56.08% across the entire analysis period. The relatively low spatially averaged coverage of 41.55% occurs during June-July-August (JJA) possibly due to more significant presence of the cloud cover during the summer monsoon.

The modeled grid cells are categorized into snow-covered, and snow-free conditions based on the simulated snow cover fraction and SWE amount. If the model derived and the corresponding MODIS derived snow cover observations agree with each other, no updates occur. If the model indicates a snow-covered grid cell, but the observation indicates snow-free condition, both SWE and snow depth states are reduced to zeros. If the model indicates a snow-free grid cell, but the observation indicates snow-covered condition, the modeled SWE during the analysis update step is increased to 5 mm, the snow

depth is increased to 0.02 m accordingly, and one layer of snowpack is created to initiate the snowpack growth. The selection of initial SWE and snow depth values in this study is completely based on Rodell and Houser (2004) and Arsenault et al. (2013). However, it is believed that a more careful sensitivity analysis should be conducted in the future to determine the optimal value to initiate the snow pack growth because adding even a thin layer of snow can have important consequences for the simulated energy balance (Rodell & Houser, 2004). In general, snow cover DA occurs more often in December-January-February (DJF) and March-April-May (MAM) (relative to JJA and September-October-November (SON)) during the major snow seasons (not shown). Details regarding Noah-MP snow parameterization, snow cover assimilation procedure, MOD10A1 preprocessing procedures, and sample snow cover DA results are described in detail in Xue et al. (2019). Details regarding how the rule-based snow cover DA could impact hydrological and energy budgets can be seen from Arsenault et al. (2013).

Xue et al. (2019) also presents the freeze/thaw assimilation procedure and sample results based on previous studies carried out by Reichle, Kumar, Mahanama, Koster, and Liu (2010) and Farhadi, Reichle, De Lannoy, and Kimball (2015). Similarly, the satellite-based freeze/thaw product used in this study is obtained from the Making Earth System Data Records for Use in Research Environments (MEaSUREs) Northern Hemisphere Polar Equal-Area Scalable Earth Grid 2.0 Daily 6 km Land Freeze/Thaw Status from the AMSR-E and the AMSR-2 (version 1; Kim, Kimball, Glassy, and Du (2017); Kim, Kimball, Glassy, and McDonald (2018)). Both morning (AM) and afternoon (PM), binary freeze/thaw states reprocessed from MEaSUREs, are employed in this study. Zeros representing the frozen landscape, and ones representing the non-frozen (or thawed) landscape. Direct updates in the top-layer soil temperature state take place twice a day at 01:30 and 13:30, which corresponds to the AM and PM MEaSUREs freeze/thaw observations, respectively, through the DA update step. The rule-based freeze/thaw assimilation scheme is conducted via comparing the agreement (or disagreement) between Noah-MP model derived and satellite-derived freeze/thaw states. To be more specific, if the model derived and the corresponding freeze/thaw observations agree with each other, no updates occur. On the other hand, if the model derived and the corresponding freeze/thaw observations do not agree, increments are computed from the surface temperature state relative to the lower surface temperature boundary and the upper surface temperature boundary of the freeze/thaw state accordingly. These increments are applied directly onto

the top-layer of soil temperature. The energy budget is mostly maintained throughout the simulation with minimal errors (not shown). Details regarding Noah-MP temperature parameterization, freeze/thaw assimilation procedure, uncertainty associated with the freeze/thaw DA procedure, and sample freeze/thaw DA results can be referred to in Xue et al. (2019).

The snow cover and freeze/thaw DA are integrated simultaneously within the dual DA scheme on a daily basis. These two updates take place individually at different time points as mentioned above. We prefer performing the snow cover DA prior to the freeze/thaw DA mainly because we expect the snow cover DA derived estimate can provide a more accurate characterization of the current snow conditions. The follow-up freeze/thaw DA can take advantage of the more accurate snow estimates, and perform temperature related updates when necessary. It is worth noting that model grid cells covered with significant amount of snowpack (i.e., greater than 50% of the snow cover fraction or greater than 5 cm of the snow depth as simulated by the Noah-MP model) are not being updated during the freeze/thaw DA due to the limited penetration depth of the 36 GHz brightness temperature channel used in the MEaSUREs detection algorithm. Therefore, freeze/thaw DA occurs much less frequently in DJF (not shown).

2.2 Study domain and models

As in Xue et al. (2021), the analysis of Part II are carried out across the HMA region bounded between 20°N and 41°N and 66°E and 101°E for a 14-year time period (2003–2016). Three sets of Noah-MP derived model simulations driven by different meteorological inputs, with and without dual assimilation scheme enabled are conducted at a spatial resolution of 0.01° ($\sim$ 1-km). Thus, a total of six sets of model simulations are evaluated in this study, which are summarized in Table 1. For example, the "HMA-GMU" experiment denotes the simulation without DA, forced by all meteorological inputs downscaled via physically-based and statistically-based George Mason University (GMU) developed algorithms onto the 0.01° grid (see Xue et al. (2021)). The "DA-HMA-GMU" (i.e., with the "DA" prefix) is the experiment forced by the same meteorological forcings as "HMA-GMU" but with both snow cover and freeze/thaw assimilation schemes enabled. Same rules applies to other four experiments, including "HMA-CHIRPS" along with its DA counterpart of "DA-HMA-CHIRPS" experiment, as well as "HMA-corr-CHIRPS" along with its DA counterpart of "DA-HMA-corr-CHIRPS" experiment.

2.3 Evaluation methods and statistics

All six experiments listed in Table 1 are integrated forward in time at a time step of 15 minutes, and the daily-averaged model output are generated. The evaluation period is from 01 February 2003 to 30 November 2016. Evaluations are conducted at three different spatial scales (i.e., point-scale, basin-scale, and domain-scale). All ground-based stations, basins, and domain extent can be seen from Xue et al. (2021). Same quality control procedures are conducted to optimize the quality of the dataset, e.g., stations (or grid cells) with records less than 200 days are excluded from the evaluation. Further, if the relative elevation difference between the 1-km scale grid cell and colocated station is greater than 50% (with the ground-based station being the baseline), we deem that the station is unrepresentative of the large-scale model estimates, and thus such stations are removed from the evaluation. It is worth noting here that although we tried some first-order criteria mentioned above to alleviate stations' under-representativeness issue, the significant disparity in horizontal resolution between model estimates and ground-based observations should not be ignored.

2.3.1 Point-scale evaluations

The performance of surface net shortwave radiation, surface net longwave radiation, skin temperature, snow depth, and near-surface soil temperatures are evaluated at daily time scales via comparisons against in-situ measurements taken by the closest colocated ground-based stations. These ground-based stations are obtained from the Coordinated Enhanced Observing Period (CEOP) Asia Monsoon project (`https://www.eol.ucar.edu/projects/ceop/dm/insitu/sites/ceop_ap/`), the Contribution to High Asia Runoff from Ice and Snow (CHARIS) project (`http://himatmap.apps.nsidc.org/hma_insitu.html`), the Global Summary of the Day (GSOD; `https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod`), the Chinese Meteorological Administration (CMA), namely the Dataset of Daily Climate Data From Chinese Surface Stations for Global Exchange (V3.0) (`http://101.200.76.197/en/?r=data/detail&dataCode=SURF_CLI_CHN_MUL_DAY_CES_V3.0`), the Central Tibetan Plateau Soil Moisture and Temperature Monitoring Network (CTP-SMTMN; Yang et al. (2013)), and the Southeastern Tibet Observation and Research Station for the Alpine Environment (SETORS; `https://data.tpdc.ac.cn/en/data/49ac37ac-0fc3-460f-83c4-c44744205474/`).

Goodness-of-fit statistics, which include bias, root mean squared error (RMSE), unbiased root mean squared error (ubRMSE), and correlation coefficient (R), are computed. Further, the level of improvement (or degradation) in DA enabled simulations is also computed to demonstrate to what extent does DA improve or degrade non-DA derived estimates in terms of all goodness-of-fit statistics. The level of improvement (or degradation) is calculated as the relative change in the absolute values of the metric obtained from the experiment and its DA counterpart. Using the bias as an example, the level of improvement (degradation), LEVEL, is calculated as:

$$LEVEL = \frac{|bias_{non-DA}| - |bias_{DA}|}{|bias_{non-DA}|}, \tag{1}$$

where the $|\cdot|$ denotes taking the absolute value of each corresponding bias value obtained from the experiment (i.e., with the "non-DA" subscript) and its DA-enabled counterpart (i.e., with the "DA" subscript). Positive LEVEL values indicate that DA-derived estimates are better, and negative values indicate that DA derived estimates are worse than non-DA derived estimates.

2.3.2 Basin-scale evaluations

The basin-scale evaluations are conducted for modeled runoff (i.e., not the routed streamflow) through comparisons against ground-based discharge measurements at the monthly scale. We do not implement any routing modules because routing related parameters at 1-km are not available at the time. Same as Xue et al. (2021), for each of the model simulation, the modeled basin-scale total runoff is computed by integrating the runoff output at each grid cell across each of the drainage basin. Figure 2 shows the five gauged basins in the study area. The ground-based runoff measurements are obtained from the Contribution to High Asia Runoff from Ice and Snow (CHARIS) project, the Department of Hydrology and Meteorology in Nepal, and the Global Runoff Data Centre, 56068 Koblenz, Germany (`https://www.bafg.de/GRDC/EN/01_GRDC/grdc_node.html`). The goodness-of-fit statistics plus the Nash–Sutcliffe model efficiency coefficient (NSE) are computed to evaluate the modeled runoff performance. Further, we compute flow duration curves for all experiments. The flow duration curve is a plot of total runoff (at the basin outlet) vs. percent of time that a particular runoff value is equaled or exceeded. In the assessment of flow duration curve agreement, following Yilmaz, Gupta, and Wagener (2008), three relative bias related statistics are computed to characterize the relative differences in the curves obtained from the model simulations and the measurements.

First, $rbias_{FMS}$ is computed to measure the relative bias in the slope of the curve for mid-flow segment (with exceedance probability between 20% and 70%), written as:

$$rbias_{FMS} = \frac{[log(x_{model,20\%}) - log(x_{model,70\%})] - [log(x_{meas,20\%}) - log(x_{meas,70\%})]}{log(x_{meas,20\%}) - log(x_{meas,70\%})}, \quad (2)$$

where $log(\cdot)$ denotes the logarithm operator, $x_{model,20\%}$ is the model simulated flow at 20% exceedance probability, $x_{model,70\%}$ is the model simulated flow at 70% exceedance probability, $x_{meas,20\%}$ is the gauge measured flow at 20% exceedance probability, and $x_{meas,70\%}$ is the gauge measured flow at 70% exceedance probability. Second, $rbias_{FHV}$ is computed to measure the relative bias in the volume of the curve for high-flow segment (with exceedance probability between 0% and 2%), written as:

$$rbias_{FHV} = \frac{\sum_{h=1}^{H}(x_{model,h} - x_{meas,h})}{\sum_{h=1}^{H} x_{meas,h}}, \quad (3)$$

where h = 1,2, $\cdots$, H are the flow indices for flows with exceedance probabilities lower than 2%. The high-flow segment can be deemed as a measure of the basin's response to heavy precipitation/snowmelt events. Third, $rbias_{FMM}$ is computed to measure the relative bias in the median value of the flow, which can also be deemed as a measure for mid-flow segment behavior, written as:

$$rbias_{FMM} = \frac{log(x_{model,med}) - log(x_{meas,med})}{log(x_{meas,med})}, \quad (4)$$

where $x_{model,med}$ is the median value of the model simulated flow, and $x_{meas,med}$ is the median value of the gauge measured flow.

2.3.3 Domain-scale evaluations

The domain-scale evaluations are conducted between model estimates and reference satellite-based products. That is, the performance of regional model output of skin temperature, and SWE are evaluated at daily time scales via comparisons against reference remotely-sensed products using the goodness-of-fit statistics. In terms of the SWE related evaluation, the satellite-based product utilized here is the Copernicus Global Land Service (CGLS) SWE product (v1.0.2; `https://land.copernicus.eu/global/products/swe`) at a spatial resolution of 5-km (Pulliainen, 2006; Takala et al., 2011) available from 01 January 2006. It provides SWE estimates between latitudes 35°N and 85°N. All model derived SWE estimates are aggregated from 0.01° onto the same 5-km CGLS SWE grid in this set of evaluation. SWE estimates in June, July, and August are excluded from evaluation due to minimized coverage of snow in summertime. Note that the CGLS SWE

product only covers about 16.3% of the entire HMA study domain, and further, mountainous regions and glaciers are excluded from the production stage.

In terms of the skin temperature related domain-scale evaluation, the reference satellite-based surface temperature products utilized here are the MODIS/Terra Land Surface Temperature Daily L3 Global 1-km Grid (MOD11A1, version 6; Wan, Hook, and Hulley (2015)) and the MODIS/Aqua Land Surface Temperature Daily L3 Global 1-km Grid (MYD11A1, version 6; Wan et al. (2015)) from 2003 to 2016. All model derived skin temperature estimates are re-gridded onto the same 1-km MODIS grid in this set of evaluation. The simple arithmetic mean of both nighttime and daytime land surface maps generated by MOD11A1 and MYD11A1 are computed as the reference satellite-based skin temperature measurements.

3 Results

3.1 Point-scale evaluations

Figure 3 shows the box plots of goodness-of-fit statistics computed from all experiments in the point-scale evaluation against ground-based measurements. In the evaluation against eight ground-based CEOP stations measuring net shortwave radiation, calculated as incoming-minus-outgoing shortwave fluxes, DA-HMA-corr-CHIRPS yields the best performance in general. In terms of median of each set of the goodness-of-fit statistics, the improvements due to DA are mostly marginal (i.e., being less than 5%). On the other hand, all experiments with dual DA enabled perform slightly better than their non-DA counterparts in terms of average RMSE, ubRMSE, and R statistics, but less so with respect to average bias. For example, compared with HMA-corr-CHIRPS, the average RMSE in DA-HMA-corr-CHIRPS improves by 13% from 64.41 W/m^2 to 56.34 W/m^2, the average ubRMSE in DA-HMA-corr-CHIRPS improves by 8% from 55.37 W/m^2 to 51.02 W/m^2, and the average R in DA-HMA-corr-CHIRPS improves by 9% from 0.55 to 0.60. It is worth noting that the statistical significance in each set of the metric difference is also tested. Unfortunately, due to the relatively small sample size and the relatively large sample variance, no statistical significance in metric difference can be claimed here. The most notable difference among DA and their non-DA counterparts can be seen from the interquartile range (IQR), calculated as the difference between the third quartile and the first quartile for each set of the goodness-of-fit statistics. The lower the IQR

is, the lower the spread is, and the higher the precision is achieved by the corresponding experiment. For example, compared with HMA-corr-CHIRPS, IQR in DA-HMA-corr-CHIRPS is reduced by 38%, 23%, 42%, and 66% for bias, RMSE, ubMRSE, and R. In addition, the improvements seen in all DA derived net shortwave radiation mostly attribute to the adjustment of snow amount during the snow cover DA. Using the HMA-corr-CHIRPS and DA-HMA-corr-CHIRPS pair as an example, the average biases are -7.44 W/m^2 and 4.33 W/m^2, respectively. The negative bias seen in HMA-corr-CHIRPS is likely due to the introduction of more precipitation, and hence more snow formation, which yields an increase in surface albedo leading to a slight increase in the outgoing shortwave radiation. The reverse in the sign of the average bias between HMA-corr-CHIRPS and DA-HMA-corr-CHIRPS pair is likely due to the reduction in the DA derived snow mass, which results in a decrease in the surface albedo, and further a slight decrease in the outgoing portion of the shortwave radiation.

In the evaluation against seven ground-based CEOP stations measuring net longwave radiation, calculated as incoming-minus-outgoing longwave fluxes, all experiments present similar performance in terms of average RMSE ($\sim$ 42 W/m^2), ubRMSE ($\sim$ 28 W/m^2), and R ($\sim$ 0.63), as well as in terms of the median of each set of the goodness-of-fit statistics. Comparatively, the performance in terms of average bias is slightly different among models where experiments forced by bias-corrected precipitation (e.g., HMA-corr-CHIRPS) tends to yield a less negative average bias (= -21.38 W/m^2) given all model derived average biases are negative values. This can be explained by the fact that more precipitation is associated with more chances of evapotranspiration, which may lead to reduction in the land surface temperature, and further result in a reduction in the outgoing portion of the longwave radiation. In addition, the IQR of bias is notably different among all DA and their non-DA counterparts. That is, in general, the reduction in IQR of bias for the experiment with dual-DA enabled is between 17% and 20%, which means DA-enabled experiments yield higher precision relative to their non-DA counterparts in terms of bias.

The snow depth evaluations shown in Figure 3 are conducted by comparing against three CHARIS stations, six CEOP stations, and eight GSOD stations. Overall, DA-HMA-CHIRPS shows slightly better performance in snow depth estimates among all experiments. Since Figure 3 is too small to visualize, using the CEOP snow depth evaluation as an example, we summarize all statistics into Table 2. It is expected that more evi-

dent differences are witnessed in HMA-corr-CHIRPS and DA-HMA-corr-CHIRPS pairs in terms of both mean and median. For example, relative to HMA-corr-CHIRPS, the median of bias in DA-HMA-corr-CHIRPS is reduced by 33% from 0.003 m to 0.002 m, the median of RMSE (ubRMSE) is reduced by 50% from 0.02 m to 0.01 m, and the median of R is improved by 65% from 0.17 to 0.28. In terms of the mean of each goodness-of-fit statistics, improvements seen with DA-enabled experiments are even higher. Although we see improvements in DA-derived snow depth estimates, again, no statistical significance of the difference can be claimed here due to the small sample size and large sample variance. In general, the agreement between model estimates and ground-based snow measurements is relatively low. For example, in the evaluation against CEOP snow depth measurements, the average correlation coefficients computed from HMA-GMU, HMA-CHIRPS, HMA-corr-CHIRPS, DA-HMA-GMU, DA-HMA-CHIRPS, and DA-HMA-corr-CHIRPS are 0.27, 0.31, 0.24, 0.29, 0.31, and 0.30. The relatively poor agreement can be attributed to erroneous model estimate itself or under-representative, erroneous, or intermittent low-quality ground-based snow measurements.

The skin temperature evaluations shown in Figure 3 are conducted by comparing against 24 CMA stations, and 11 CEOP stations. All model derived estimates show relatively low absolute bias (< 0.15 K) and high correlation coefficient (~ 0.97), especially in the evaluation against CMA skin temperature measurements. In general, DA-HMA-GMU and DA-HMA-CHIRPS demonstrate slightly better performance in skin temperature estimates among all experiments, especially in RMSE, ubRMSE, and R statistics, but less so with respect to the bias. It is encouraging to see that all DA-derived estimates tend to improve the estimates relative to their non-DA counterparts, notably in HMA-corr-CHIRPS and DA-HMA-corr-CHIRPS pairs. For example, in the evaluation against CMA skin temperature measurements, relative to HMA-corr-CHIRPS, DA-HMA-corr-CHIRPS improves the average RMSE (ubRMSE) by 9% (10%) from 3.71 K (2.85 K) to 3.36 K (2.57 K).

Figure 4 further summarizes the relative improvement computed from each set of DA derived estimate relative to non-DA derived ones in the evaluation against ground-based measurements as a function of seasonality in terms of all goodness-of-fit statistics. As stated above, the majority of the improvements seen in DA are mostly for random errors reductions, such as RMSE and ubRMSE across all seasons. The magnitudes of the improvement in RMSE and ubRMSE are typically smaller during June-July-August-

September due to least amounts of disagreement seen between model derived estimates and satellite-based observations for both snow cover and freeze/thaw states. In terms of bias, the results are mixed across different seasons. For example, in the evaluation against CEOP net shortwave radiation, all DA-derived estimates perform slightly better through June and November, whereas perform worse through December and May relative to non-DA counterparts. Using the HMA-CHIRPS and DA-HMA-CHIRPS pair as an example, HMA-CHIRPS yields average biases of -14.05 W/m^2 during JJA and -6.33 W/m^2 during SON, which are both negative. DA-HMA-CHIRPS improves the negative bias slightly, and yields average biases of -12.90 W/m^2 during JJA and -3.85 W/m^2 during SON. However, during DJF and MAM, the average biases computed from HMA-CHIRPS (DA-HMA-CHIRPS) are positive, i.e., 5.17 W/m^2 (18.42 W/m^2) during DJF and 18.62 W/m^2 (23.20 W/m^2) during MAM. The exact reason for the average bias sign change issue across different seasons remain unclear. The errors may arise from many sources, e.g., 1) relatively large in-situ measurement errors during wintertime when snow is present, and/or 2) inaccurate representation of Noah-MP model physics during wintertime. Further, during DJF and MAM, we see DA-HMA-CHIRPS exacerbates the positive bias seen from HMA-CHIRPS. Without accurate ground-based snow measurements at all colocated stations, it is difficult to discern the reasons clearly. It is plausible that during JJA and SON, which likely contain the time periods of the end and start of the snow seasons, the rule-based snow cover DA is beneficial at capturing quick transitions between snow-on and snow-off conditions. However, during DJF and MAM, sometimes, the Noah-MP modeled snow melts too slowly without DA but being removed too quickly when using rule-based DA (see Figure 7, CEOP Station#2 for an example)). As discussed above, the surface net shortwave radiation is more dictated by the surface property. The overall reduction in the snow cover due to DA tends to decrease outgoing shortwave radiation, and hence, results in an increase in the net shortwave radiation, which leads to an exacerbation of the positive biases seen between December and May and an improvement of the negative biases seen between June and November.

In terms of R, the improvements/degradations are typically negligible across all seasons except for snow depth evaluations. For example, we see relatively high level of improvements in DA-HMA-corr-CHIPRS derived snow depth in terms of R as well as the other three goodness-of-fit statistics across all seasons (i.e., excluding summer seasons). These improvements (i.e., reduction in snow depth) mostly benefit from the systematic

reduction of the over-correction issue seen in the bias-corrected CHIRPS through the snow cover DA, which directly adjusts the inter-annual variability of the snow cover, and hence mitigate the negative effects brought by the fixed long-term precipitation correction factors.

Figure 5 shows the box plots of goodness-of-fit statistics computed from all experiments in the point-scale evaluation against ground-based soil temperature measurements. All model derived soil temperature estimates generally present relatively good agreement with all sources of soil temperature measurements, e.g., the average Rs are all greater than 0.9. In the evaluation against 63 CTP-SMTMN 0-5 cm soil temperature stations, although all DA-derived estimates tend to improve the estimates obtained from their non-DA counterparts in terms of all goodness-of-fit statistics, the improvements are marginal (i.e., less than 1%). The marginal improvements are expected partly because of the relatively good agreement seen between satellite-based freeze/thaw observations and Noah-MP simulated estimates outside the Hindu-Kush Karakoram Himalaya region. Hence, a relatively small number of analysis updates in the freeze/thaw DA is not expected to impact the overall magnitude of improvements/degradations significantly (Xue et al., 2019). In addition, the non-significant improvements seen in the evaluation against CTP-SMTMN may be partly attributed to the relatively low spatial variability in soil temperature measurements in spite of the relatively large number of ground-based stations as discussed in Xue et al. (2021). Similar marginal improvements in DA-derived estimates in terms of the mean of each goodness-of-fit statistics are also seen in the evaluation against CEOP 4 cm soil temperature, SETORS 4 cm soil temperature, and CEOP 5 cm soil temperature.

On the other hand, we see marginal degradations in the DA-derived 3 cm soil temperature estimates relative to non-DA counterparts in the evaluation against one CEOP station. Using the bias as an example, except for HMA-corr-CHIRPS, the other two sets of DA derived estimates move to the wrong direction which further exacerbate the positive bias seen in the non-DA derived estimates. We do not see the same issue in 4 cm and 5 cm soil temperature evaluations because the non-DA derived estimates all yield negative biases. However, in HMA-corr-CHIRPS with an overly-corrected precipitation (i.e., without considering the inter-annual variability in precipitation correction factors), which yields a slight negative bias in the 3 cm soil temperature likely due to the evaporative cooling effect, DA-corr-CHIRPS moves towards the right direction and yields a

close to zero bias. It is difficult to discern exactly which factor may result in such discrepancies in DA performance as well as the bias sign change issue among different layers of the soil. Since there is only one CEOP station measuring the 3 cm soil temperature, the measurement errors are difficult to avoid. It is also possible that the relatively simple linear interpolation method used to apply with the modeled soil temperature estimates to match with the measurement depth may not be appropriate in this case because the temperature gradient may not be linear. Similarly, we further summarizes the relative improvement computed from each set of DA derived estimate relative to non-DA derived ones in the evaluation against ground-based soil temperature measurements as a function of seasonality in terms of all goodness-of-fit statistics in Figure 6. In general, we see relatively high magnitude of DA improvements/degradations across DJF and MAM seasons, rather than in JJA and SON. Although the majority of improvements are seen in DA-derived estimates (i.e., the maximum level of improvements are $\sim 40\%$), not surprisingly, we still see degradations mainly taking place in the evaluation against CEOP 3 cm soil temperature especially in terms of bias. Therefore, Figure 6 further corroborates the findings concluded in Figure 5.

Figure 7 shows example time series of several model output for CEOP station#1 at 31.37°N, 91.90°E from 01 August 2003 to 01 October 2003, and for CEOP station#2 at 31.93°N, 91.71°E from 01 February 2007 to 01 April 2007. The two stations are selected because they are among the very few stations measuring multiple variables at the same time. The two time periods are selected to highlight the impact of dual DA across different seasons. In addition, the benefits of the GMU developed downscaling scheme can also been seen from this Figure. For example, in between 01 August 2003 and 01 October 2003, when snow starts to accumulate, experiments with GMU downscaled precipitation, relative to other two CHIRPS precipitation variants, are more capable of catching the snow depth spike shown in the CEOP measurement on 04 September 2003. With a more accurate snow accumulation, both HMA-GMU and DA-HMA-GMU derived skin temperature estimates on 04 September 2003 align more closely with the CEOP measurement.

Between 01 August 2003 and 01 September 2003, no snow is present. Satellite-derived snow cover generally agrees with model derived snow cover for this location as snow-free. Therefore, snow cover DA has almost zero effect in net shortwave radiation and snow depth estimates. In this case, freeze/thaw DA mainly controls the DA-derived estimates

in terms of net longwave radiation (i.e., partly modulated by skin temperature), skin temperature, and soil temperature profile. The maximum difference seen in all DA-derived 4 cm soil temperature relative to their non-DA counterparts are $\sim$ -1.2K. As all non-DA derived soil temperature already yields relatively high negative biases as compared with CEOP measurements, a further decrease in the soil temperature as a result of the freeze/thaw DA implementation moves the estimate to the wrong direction in terms of bias. However, it is encouraging to see the temporal variability in the estimates are better picked up by the DA-derived estimates. Using the HMA-GMU and DA-HMA-GMU as an example, between 18 August 2003 and 19 August 2003, CEOP measured 4 cm soil temperature witnesses a daily temperature drop of $\sim$ 3K (from 287.94 K to 284.93 K). Without DA, HMA-GMU is only able to model the drop as $\sim$ 0.6 K (from 281.17 K to 280.55 K). With freeze/thaw DA, DA-HMA-GMU models the drop as $\sim$ 1.5 K (from 281.10 K to 279.59 K). Similar improvements in the temporal variability agreement are also witnessed in the other DA-derived experiments as well as during other time periods. We acknowledge that such improvements in modeling soil temperature temporal variability are still far from being accurate as compared with the ground-based measurements. These discrepancies may be attributed to the errors in the ground-based measurement itself or in the model errors, e.g., from the relatively simple and conservative strategy that we used in applying the temperature increments during the freeze/thaw DA analysis step (see Xue et al. (2019) for discussions).

Between 01 February 2007 and 01 April 2007, we see a typical and complete snow accumulation and ablation time series measured by CEOP in Figure 7h. During the snow accumulation phase from 01 February 2007 to 14 March 2007, it is not surprising to see that snow depth estimates derived by HMA-corr-CHIRPS are almost three to four times greater than the measurement. With the rule-based snow cover DA, DA-HMA-corr-CHIRPS is able to get rid of approximately half of the snow mass. Comparatively, DA-HMA-CHIRPS yield the best agreement with all ground-based measurements during the accumulation phase. After 14 March 2007, all DA derived estimates tend to melt down all snow pack within two days to match with the satellite-based snow cover flag, while all non-DA experiments take a longer period to melt down all snow pack, especially for HMA-GMU and HMA-CHIRPS experiments. The difference in the snow melting speed significantly affects surface energy budget as snow cover dramatically regulate the amount of energy being reflected from the surface. With more snow accumulating on the ground during

the melting phase, HMA-GMU and HMA-CHIRPS yield higher outgoing shortwave radiation, which results in lower net shortwave radiation values. Further, due to the decrease in the snow surface temperature, HMA-GMU and HMA-CHIRPS yield relatively low outgoing longwave radiation, which result in less negative net longwave radiation. Overall, HMA-CHIRPS yields the best agreement with all ground-based measurements during the ablation phase. DA derived estimates are not superior during the snow melt phase probably because they melt the snow pack more quickly than it should be. This phenomenon is caused by the simple snow mass reduction rule given by the snow cover DA algorithm (see Section 2.1).

3.2 Basin-scale evaluations

Figure 8 shows the flow duration curves calculated from all model simulations as well as ground-based measurements for the five gauged basins from 2003 to 2016. We partition each of the curves into high-flow segment (with exceedance probability between 0% and 2%), high-to-mid flow segment (with exceedance probability between 2% and 20%), mid-flow segment (with exceedance probability between 20% and 70%), and low-flow segment (with exceedance probability between 70% and 100%), following Yilmaz et al. (2008). We choose to show flow duration curves on the log scale rather than the total runoff time series on the normal scale because it is rather difficult to clearly visualize different model simulations using the total runoff time series on the normal scale. Overall, it is not surprising that none of the experiments can perfectly reproduce the flow duration curve calculated from ground-based measurements, possibly due to 1) measurement errors related to human impacts, and/or 2) model errors arising from imperfect forcings and model structures.

According to the glacier map obtained from the Global Land Ice Measurements from Space (GLIMS) project (GLIMS & NSIDC, 2005) at a spatial resolution of 0.01°, Basins #1 and #5 are both with glaciated fraction of 0% and experience negligible seasonal snow cover (i.e., maximum snow coverage being less than 10% in the evaluation period according to the MOD10A1 snow cover product). It is expected that relatively small differences exist between different model simulations. In other words, model simulations with and without the dual DA assimilation scheme present comparable performance in reproducing the flow duration curves derived from ground-based measurements as shown in Figure 8. On the other hand, Basins #2 through #4 are snowmelt fed (or snow-glacier-melt

fed) basins. As discussed above, snow cover DA significantly impacts snow melt timing, which further impacts the runoff modeling performance. For example, HMA-corr-CHIRPS yields significantly higher magnitude in the high-flow segment mainly due to the overly-corrected precipitation forcings, which is likely limited by the fixed long-term correction factors without considering the inter-annual variability of the total precipitation. With snow cover DA, DA-HMA-corr-CHIRPS improves the $rbias_{FHV}$ significantly across all three basins, especially for Basin #4 (see Table 3). Improvements in the high-flow segments as well as in the high-to-mid flow segments are also seen in other DA-derived estimates across all three basins. Further, we compute both $rbias_{FMS}$ and $rbias_{FMM}$ statistics to quantify the systematic errors in the model simulated flow at mid-flow segment. $rbias_{FMS}$ focuses on the slope of the curve whereas $rbias_{FMM}$ focuses on the magnitude of the median value. Although the majority of the DA-derived estimates show improvements in the mid-flow segment, slight degradations notably in the $rbias_{FMS}$ are witnessed. These degradations in DA derived $rbias_{FMS}$ may arise from 1) the lack of river routing scheme in the model, and 2) the inaccuracy in updating snow melt rate within the simple rule-based DA (see Figure 7 as an example).

In terms of the low-flow segment, which is a measure of the long-term base flow, it is encouraging to see that all DA derived estimates tend to move towards the ground-based measurements due to snow cover DA. It is worthy of noting that we still see discrepancies between measured and modeled duration curves in the low-flow segment. According to GLIMS, the glaciated fraction is 15.45% (Basin #2), 5.53% (Basin #3), and 17.8% (Basin #4). As the low-flow segment can be deemed as a measure of the long-term base flow, it is postulated that the lack of the glacier modeling routine may negatively impact the runoff simulation accuracy. This explanation is based on a separate study (not shown) during which we study the impact of including a relatively simplistic glacier modeling routine (i.e., using a relatively simple glacier land cover representation to modify the energy balance and water balance associated with the glaciated grid cells) into the Noah-MP model. It is found that the the inclusion of a simplistic glacier model could contribute to a more accurate representation of model simulated base flow component originating from snow and glacier melt. Since the Part I manuscript did not include any glacier modelings, and further, current Noah-MP does not support a full glacier mass balance model yet, all glacier related analysis are not shown here. Instead, we will demon-

strate the effect of glacier routines in streamflow modelings in a separate study in the future.

Figure 9 further summarizes the goodness-of-fit plus NSE statistics for all experiments across three snow-dominated basins, i.e., Basin #2, Basin #3, and Basin #4, in the evaluation against ground-based monthly runoff measurements from 2003 to 2016. It is encouraging to see that DA-derived estimates improve all goodness-of-fit statistics across all three basins relative to non-DA derived counterparts. In general, DA-HMA-corr-CHIRPS shows its superiority in runoff simulations over snowmelt fed (or snow-glacier-melt fed) basins relative to other experiments. For example, in Basin #2, relative to HMA-corr-CHIRPS, DA-HMA-corr-CHIRPS reduces the bias by 69% (from 88.23 m^3/s to 26.92 m^3/s), RMSE by 30% (from 154.73 m^3/s to 107.79 m^3/s), and ubRMSE by 18% (from 127.11 m^3/s to 104.37 m^3/s). DA-HMA-corr-CHIRPS also improves the R (NSE) from 0.87 (0.43) to 0.90 (0.72). These findings suggest that the improvements in snow estimates due to snow cover DA can translate into improved runoff estimates.

3.3 Domain-scale evaluations

Figure 10 shows the goodness-of-fit statistics computed for HMA-GMU and DA-HMA-GMU in the evaluation against the MODIS skin temperature product from 2003 to 2016 across HMA at a spatial resolution of 1-km. We only show the evaluation of HMA-GMU and DA-HMA-GMU in Figure 10 because the other two experiment pairs demonstrate similar patterns and findings. Although both HMA-GMU and DA-HMA-GMU present relatively good agreement with MODIS skin temperature in terms of the temporal variability (i.e., with the spatially-averaged R = 0.94), it is not too surprising to see that skin temperature magnitudes vary significantly among model estimates as well as satellite-derived reference because neither the model nor the satellite derived estimate should be treated as "truth". The worst agreement (i.e., relatively high magnitudes of bias, RMSE, ubRMSE, and low R) occurs along the Hindu-Kush Karakoram Himalaya region as discussed in Xue et al. (2021) mainly due to the complicated cryospheric-hydrological process compounded by the complex terrain. With dual DA enabled, DA-HMA-GMU tends to improve all goodness-of-fit statistics except for R, which generally results in a better agreement with the reference measurement, especially notable along the Karakoram-Himalaya region. The spatially-averaged improvements seen in DA-derived bias, RMSE, and ubRMSE are between 6% to 7% relative to the non-DA counterpart. The range of

the change (i.e., DA-HMA-GMU minus HMA-GMU) in the spatially- and temporally-averaged mean in the skin temperature is from 9.1 K per day to -4.5 K per day across all grid cells (not shown). Compared with HMA-GMU, DA-HMA-GMU increases the skin temperature estimates mainly across the Tibetan Plateau, which are likely due to the more frequent presences of freeze/thaw and snow cover status discrepancies between model estimates and satellite observations. Figure 11 further corroborates what is observed in Figure 10. It shows the box plot of change in the ubRMSE (ΔubRMSE) computed between HMA-GMU and DA-HMA-GMU. Note that we do not show box plots for other computed metrics because they are similar to what shown in Figure 11. Figure 11 is binned as a function of the elevation per grid cell. The spatially-distributed elevation per grid cell (Elev; meters) throughout the assessment period are binned into eight categories, including (1) $0 \leq \text{Elev} \leq 1000$, (2) $1000 < \text{Elev} \leq 2000$, (3) $2000 < \text{Elev} \leq 3000$, (4) $3000 < \text{Elev} \leq 4000$, (5) $4000 < \text{Elev} \leq 5000$, (6) $5000 < \text{Elev} \leq 6000$, (7) $6000 < \text{Elev} \leq 7000$, and (8) $\text{Elev} > 7000$. The sample sizes (number of grid cells) for the six bins are 3216572, 1003864, 505946, 556549, 1362142, 669144, 12317, and 278, respectively. The positive ΔubRMSE indicate skill improvements in the DA-HMA-GMU relative to HMA-GMU. Therefore, it is seen that the majority of the skin temperature improvements in terms of ubRMSE take place at grid cells with elevations between 2000 m and 7000 m. We also see slight degradations in skin temperature estimates for grid cells with elevations above 7000 m. This is expected because 94% of the grid cells with elevations above 7000 m are glacierized according to GLIMS. The degradation in DA could be attributed to 1) inaccurate retrievals of MODIS land surface temperature, and/or 2) inaccurate physical representations of snow accumulation and melt processes on top of glaciers within the model itself. In general, DA-HMA-GMU and DA-HMA-CHIRPS demonstrate the best agreement with MODIS skin temperature estimates, which corroborates the results in the ground-based skin temperature analysis.

Figure 12 shows the goodness-of-fit statistics computed for HMA-GMU and DA-HMA-GMU in the evaluation against the CGLS SWE product from 2006 to 2016 above latitude 35°N at a spatial resolution of 5-km. Again, we only show the evaluation of HMA-GMU and DA-HMA-GMU in Figure 12 because the other two experiment pairs demonstrate similar patterns and findings. Similar to the evaluations above, it is expected that SWE estimates vary significantly among model estimates as well as satellite-derived reference because neither the model nor the satellite derived estimate should be treated as

"truth". Although DA-HMA-GMU improves R slightly relative to HMA-GMU in the evaluation against the reference product, the agreement in the temporal variability is generally low (i.e., R < 0.4), which emphasizes that the snow estimation across such complex terrain remains a challenging task. The worst agreements (i.e., relatively high magnitudes of bias, RMSE, ubRMSE, and low R) are mostly colocated with relatively high elevation regions inside the Tibetan Plateau relative to the Taklamakan desert mainly due to the difference in climate regions as discussed in Xue et al. (2021). With dual DA enabled, DA-HMA-GMU tends to improve all goodness-of-fit statistics, which results in a better agreement with the reference measurement. Within the CGLS covered domain (i.e., above latitude 35°N, excluding mountainous regions), the range of the change (i.e., DA-HMA-GMU minus HMA-GMU) in the spatially- and temporally-averaged mean in the SWE is from 3 mm per day to -65 mm per day across all grid cells (not shown). It is expected that the snow reduction magnitude in DA enabled experiments are generally higher than the snow addition magnitude because the rule-based snow update algorithm is more promptly at reducing the snowpack at qualified grid cells. Overall, within the CGLS covered domain, DA-HMA-GMU adds $\sim 3\%$ more SWE relative to HMA-GMU (from 0.88 mm per day to 0.91 mm per day). Since CGLS SWE yields the spatially- and temporally-averaged mean of 2.83 mm per day, DA-HMA-GMU tends to move closer to the CGLS SWE estimates, which results in a better agreement in terms of all goodness-of-fit statistics. In general, DA-HMA-GMU and DA-HMA-CHIRPS demonstrate the best agreement with CGLS SWE estimates, which are consistent with the analysis in the point-scale snow depth evaluations.

4 Conclusions and discussions

This second paper of the two-part series focuses on demonstrating the impact of assimilating satellite-based snow cover and freeze/thaw observations into the hyper-resolution, offline terrestrial modeling system across HMA from 2003 to 2016. To this end, this study systematically evaluates a total of six sets of 0.01° ($\sim$ 1-km) model simulations forced by different precipitation forcings, with and without the dual assimilation scheme enabled, at point-scale, basin-scale, and domain-scale. Key conclusions drawn from this study are summarized below:

1) In the evaluation against ground-based net shortwave radiation measurements, DA-HMA-corr-CHIRPS yields the best performance. The DA-enabled experiment tends

to yield a lower IQR (i.e., higher precision) than their non-DA counterpart in all goodness-of-fit statistics.

2) In the evaluation against ground-based net longwave radiation measurements, all experiments present similar performance in terms of average (median) RMSE, ubRMSE, and R. The DA-enabled experiment tends to yield a lower IQR (i.e., higher precision) relative to their non-DA counterpart in terms of bias.

3) In the evaluation against ground-based snow depth measurements, DA-HMA-CHIRPS shows slightly better performance. Further, DA derived estimates are generally better than non-DA counterparts in terms of all goodness-of-fit statistics.

4) In the evaluation against ground-based skin temperature measurements, DA-HMA-GMU and DA-HMA-CHIRPS demonstrate slightly better performance among all experiments, especially in RMSE, ubRMSE, and R statistics, but less so with respect to the bias. All DA-derived estimates tend to improve the estimates relative to their non-DA counterparts.

5) In the evaluation against ground-based near-surface soil temperature profile measurements, the results are mixed. All DA-derived estimates tend to improve the estimates obtained from their non-DA counterparts in terms of all goodness-of-fit statistics, however, the improvements are marginal. These marginal improvements are mainly seen in the evaluation against CTP-SMTMN 0-5 cm soil temperature, CEOP 4 cm soil temperature, SETORS 4 cm soil temperature, and CEOP 5 cm soil temperature. On the other hand, marginal degradations are seen in the DA-derived 3 cm soil temperature estimates relative to non-DA counterparts in the evaluation against one CEOP station.

6) Across three snowmelt fed (or snow-glacier-melt fed) basins, DA-derived runoff estimates demonstrate better performance relative to non-DA derived counterparts across low-flow, high-flow, as well as high-to-mid flow segments. Overall, DA-HMA-corr-CHIRPS shows its superiority in runoff simulations relative to other experiments. These results suggest that the improvements in snow estimates due to snow cover DA can translate into improved runoff estimates.

7) In the evaluation against the MODIS skin temperature product as well as in the evaluation against the CGLS SWE product, DA-HMA-GMU and DA-HMA-CHIRPS yield the best agreement. The majority of the skin temperature improvements due to DA take

place at grid cells with elevations between 2000 m and 7000 m. Slight degradations in skin temperature estimates due to DA are witnessed for grid cells with elevations above 7000 m (i.e., mostly covered with glaciers). In terms of domain-scale SWE evaluation, SWE estimates vary significantly among model estimates as well as satellite-derived reference. The agreement in the temporal variability among different products/estimates is generally low, which emphasizes that the snow estimation across such complex terrain remains a challenging task.

8) In terms of seasonality, in general, the majority of the improvements seen in DA are mostly for random errors reductions, such as RMSE and ubRMSE across all seasons and the magnitude of improvements are typically smaller during JJA seasons.

In summary, the proposed dual-assimilation system is beneficial in improving the cryospheric-hydrological process within the Noah-MP land surface model for use in HMA. When forced with different precipitation forcings, the performance of DA-enabled experiments differs. In general, DA-HMA-corr-CHIRPS demonstrates its superiority in runoff estimates whereas DA-HMA-GMU and DA-HMA-CHIRPS demonstrate superiority in other model estimates (e.g., snow depth). The good performance obtained from model derived runoff simulations forced by the bias-corrected CHIRPS is partly expected because the precipitation input was calibrated using ground-based runoff measurements (Beck et al., 2020). It is encouraging to see that the proposed dual DA system successfully mitigates the negative effects brought by the overly-corrected precipitation forcings due to the fixed long-term correction factors without considering the inter-annual variability of the total precipitation.

Although we try hard to gather all reference measurements across HMA to evaluate, we acknowledge that the sample size used in the evaluation process is far from being adequate. Therefore, with limited evaluation references across HMA, we choose not to explicitly refer to the best DA derived experiment among DA-HMA-GMU, DA-HMA-CHIRPS, and DA-HMA-corr-CHIRPS as the "HMA-LDAS". Instead, we focus on emphasizing the effect of the proposed dual DA procedures for use across HMA.

While we are encouraged with the improvements seen in the proposed HMA-LDASs, the limitations of the current study are also worth mentioning here. These limitations should be addressed in the future studies in order to further improve HMA-LDASs performance. These limitations include, but not limited to, 1) the lack of glacier modeling

routines, 2) the lack of river routing schemes along with a detailed human land-water management representation, 3) the lack of a more accurate representation of the snow depletion process used within the simplistic rule-based snow cover DA, 4) the lack of a more detailed soil temperature profile within Noah-MP, and 5) the lack of a more accurate representation of the relationship between skin temperature change and top-layer soil temperature change when used within the simplistic rule-based freeze/thaw DA. Further, the snow cover DA mostly dominates the impact on all key variables of interest as seen from the analysis. Given the marginal improvements seen within the freeze/thaw DA, a more thoughtful rule-based strategy probably should be established to not only capture the changes in the soil temperature states, but also to better reveal the updates in the water-related variables. We acknowledge that the relatively simplistic rule-based multi-variate DA system is not a panacea, more advanced DA techniques may be used in order to facilitate the development of the HMA-LDAS as future versions (i.e., version 2, version 3, etc.). But we believe that the presented study here (i.e., two-part series) did point out useful directions towards future studies for modeling such a challenging HMA region, e.g., hyper-resolution modeling coupled with multi-variate assimilation strategies. These HMA-LDASs will be extremely useful for future studies 1) to understand surface flux, snow/ice storage, and water balance changes in HMA and investigate the causality of these changes at the regional to local scale, 2) to develop assessments of significant trends/changes in both observed (e.g. snow) and unobserved (e.g. groundwater storage change) states of land surface, and 3) to develop better forecasts of hydrological extremes such as droughts and floods.

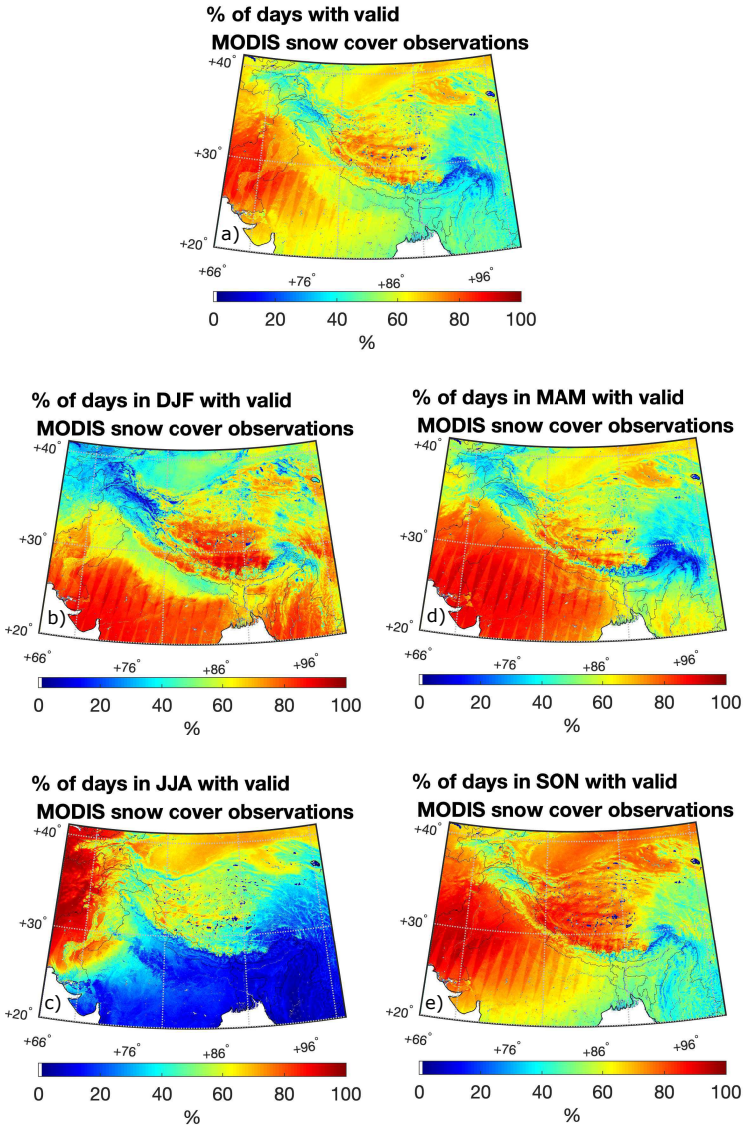

Figure 1. a) The spatial map of the total percentage of days with valid MODIS snow cover observations from 01 February 2003 to 30 November 2016. Seasonal MODIS data availability maps are shown in b) through e). DJF = December, January, February; MAM = March, April, May; JJA = June, July, August; SON = September, October, November.

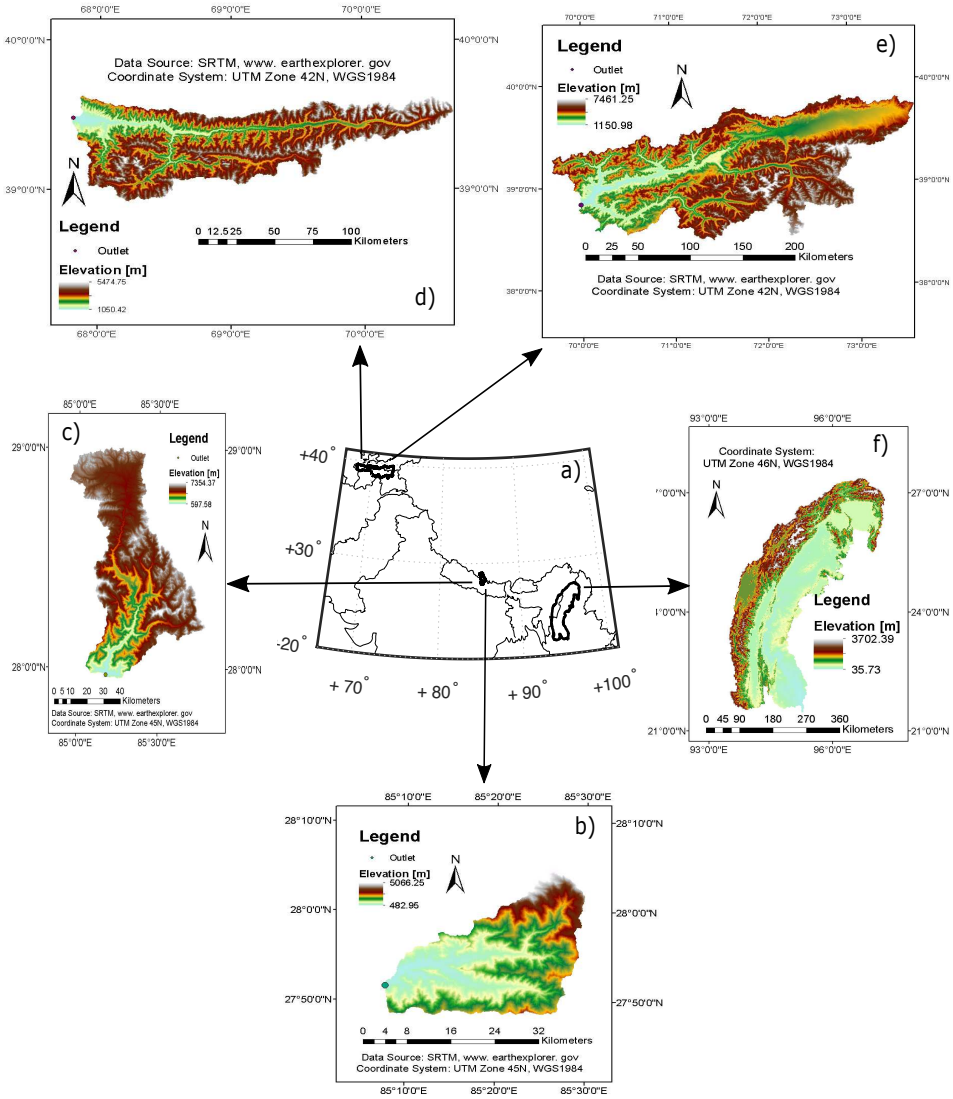

Figure 2. a) HMA study domain with gauged basin outlines in black. Gauged Basin #1 through Basin #5 are shown in b) through f) with elevation information and basin outlet locations.

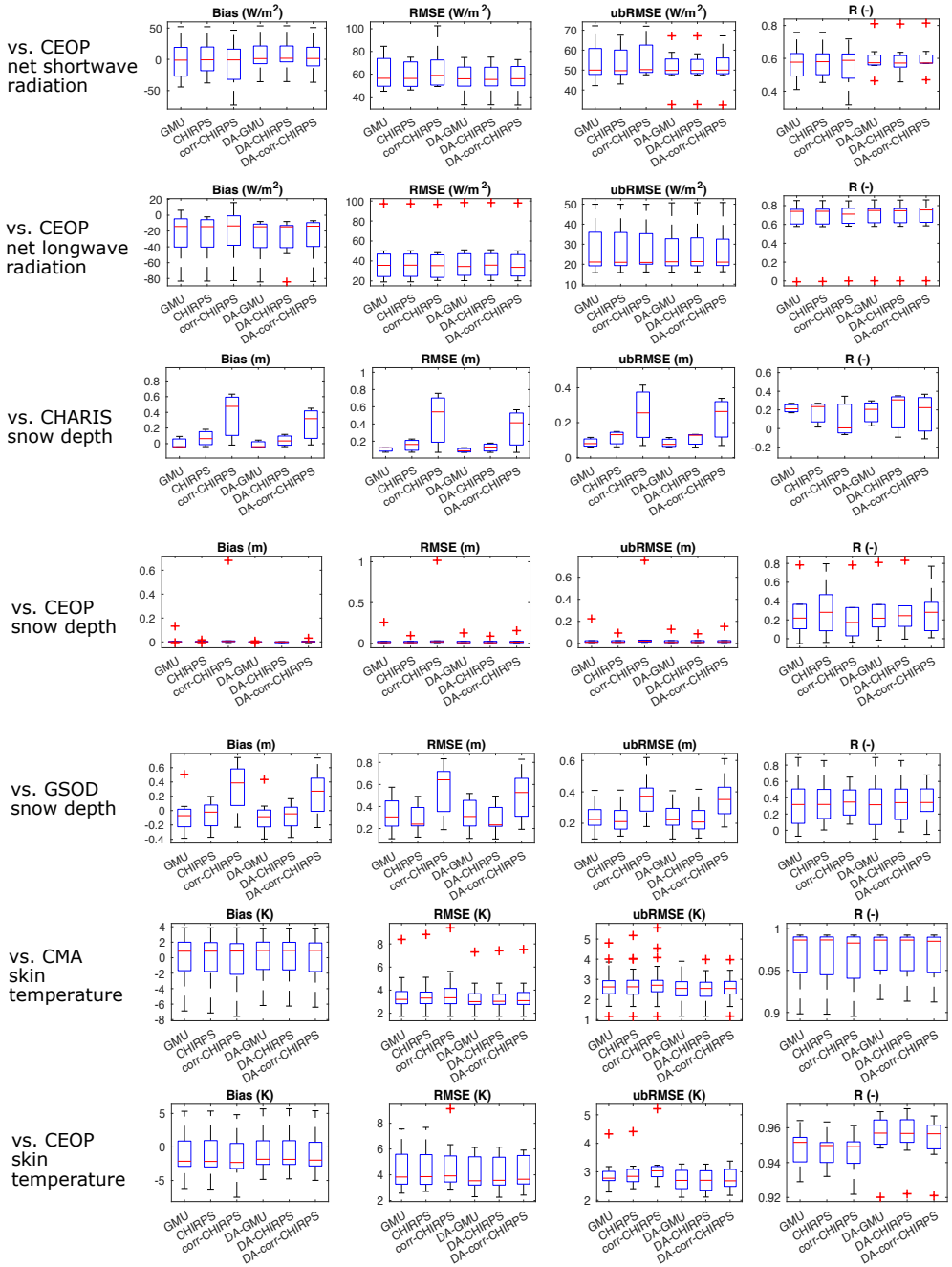

Figure 3. Box plots of bias (column 1), RMSE (column 2), ubRMSE (column 3), R (column 4) computed from HMA-GMU, HMA-CHIRPS, HMA-corr-CHIRPS, DA-HMA-GMU, DA-HMA-CHIRPS, and DA-HMA-corr-CHIRPS in the evaluation against ground-based CEOP net shortwave radiation (row 1), CEOP net longwave radiation (row 2), CHARIS snow depth (row 3), CEOP snow depth (row 4), GSOD snow depth (row 5), CMA skin temperature (row 6), and CEOP skin temperature (row 7). The common experimental name of "HMA" is omitted for clarity. The plus signs and red lines in the box plots are shown as outliers and medians, respectively. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively.

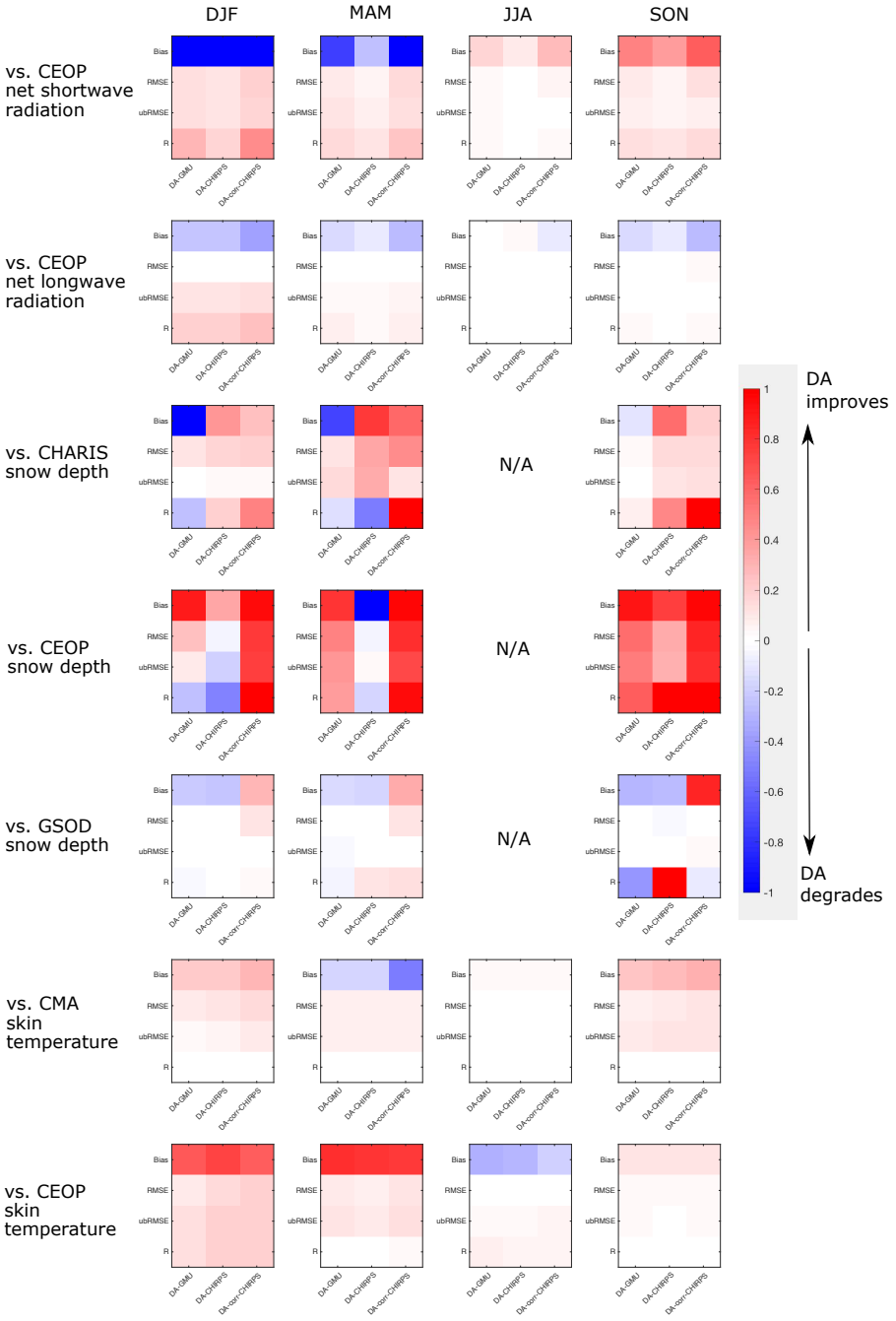

Figure 4. Level of improvement (red) or degradation (blue) in the DA-HMA-GMU (relative to HMA-GMU), DA-HMA-CHIRPS (relative to HMA-CHIRPS), and DA-HMA-corr-CHIRPS (relative to HMA-corr-CHIRPS) in the evaluation against ground-based CEOP net shortwave radiation (row 1), CEOP net longwave radiation (row 2), CHARIS snow depth (row 3), CEOP snow depth (row 4), GSOD snow depth (row 5), CMA skin temperature (row 6), and CEOP skin temperature (row 7) as a function of seasonality in terms of bias, RMSE, ubRMSE, and R. DJF = December, January, February; MAM = March, April, May; JJA = June, July, August; SON = September, October, November. All snow depth box plots during JJA are not provided due to minimized snow coverage, and hence are denoted as "N/A".

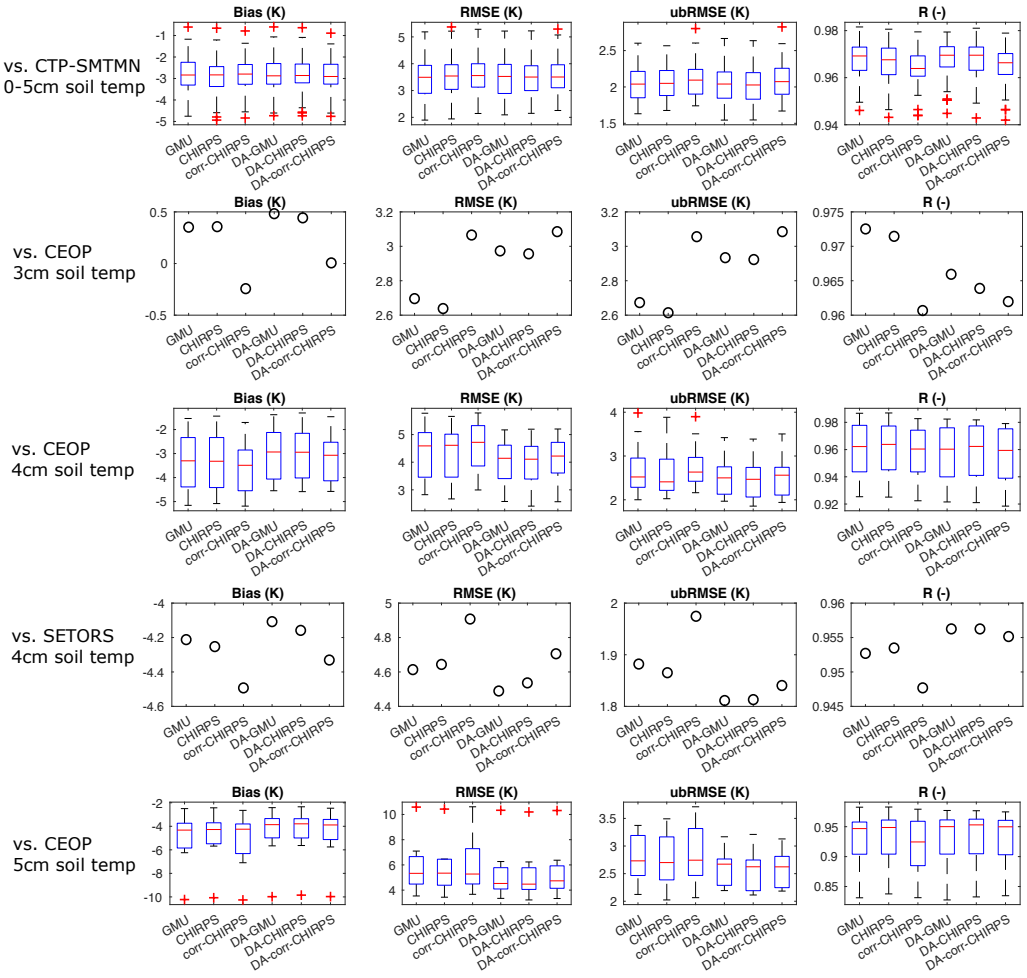

Figure 5. Same as Figure 3, but for the evaluation against ground-based CTP-SMTMN 0-5 cm soil temperature (row 1), CEOP 3 cm soil temperature (row 2), CEOP 4 cm soil temperature (row 3), SETORS 4 cm soil temperature (row 4), and CEOP 5 cm soil temperature (row 5). Note there is only one CEOP station measuring 3 cm soil temperature, and there is only one SETORS station.

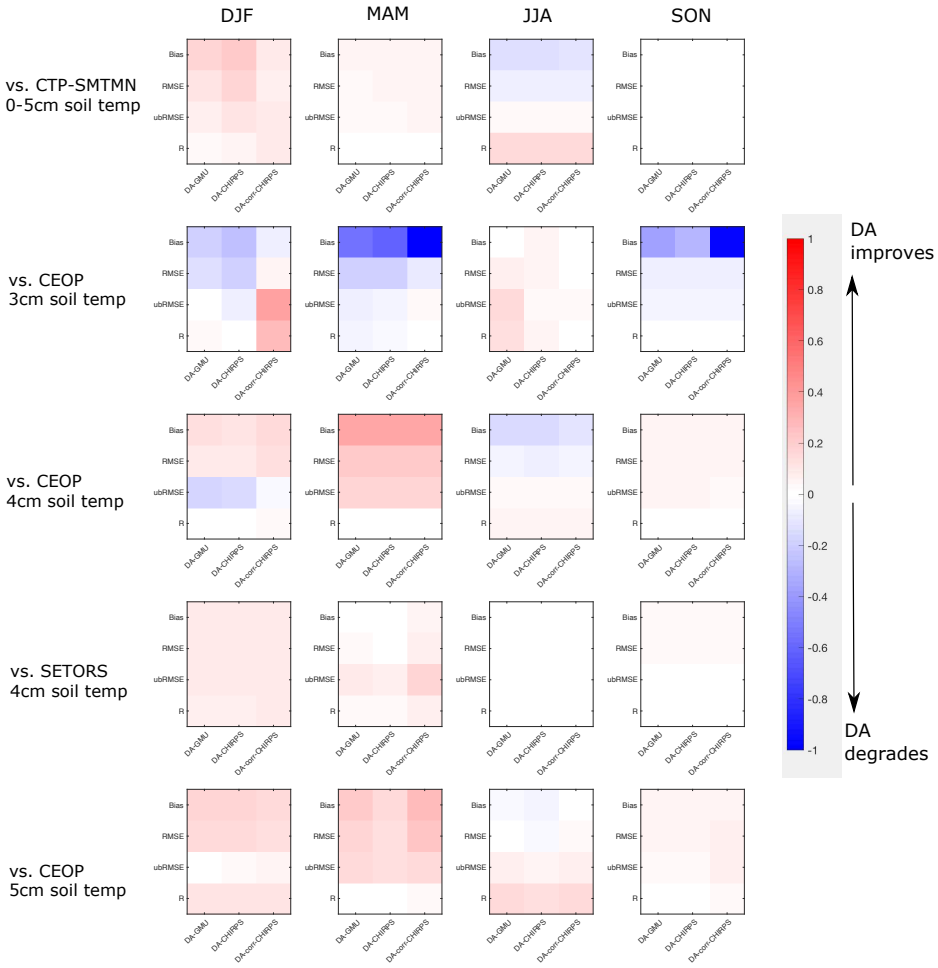

Figure 6. Same as Figure 4, but for the evaluation against ground-based CTP-SMTMN 0-5 cm soil temperature (row 1), CEOP 3 cm soil temperature (row 2), CEOP 4 cm soil temperature (row 3), SETORS 4 cm soil temperature (row 4), and CEOP 5 cm soil temperature (row 5).

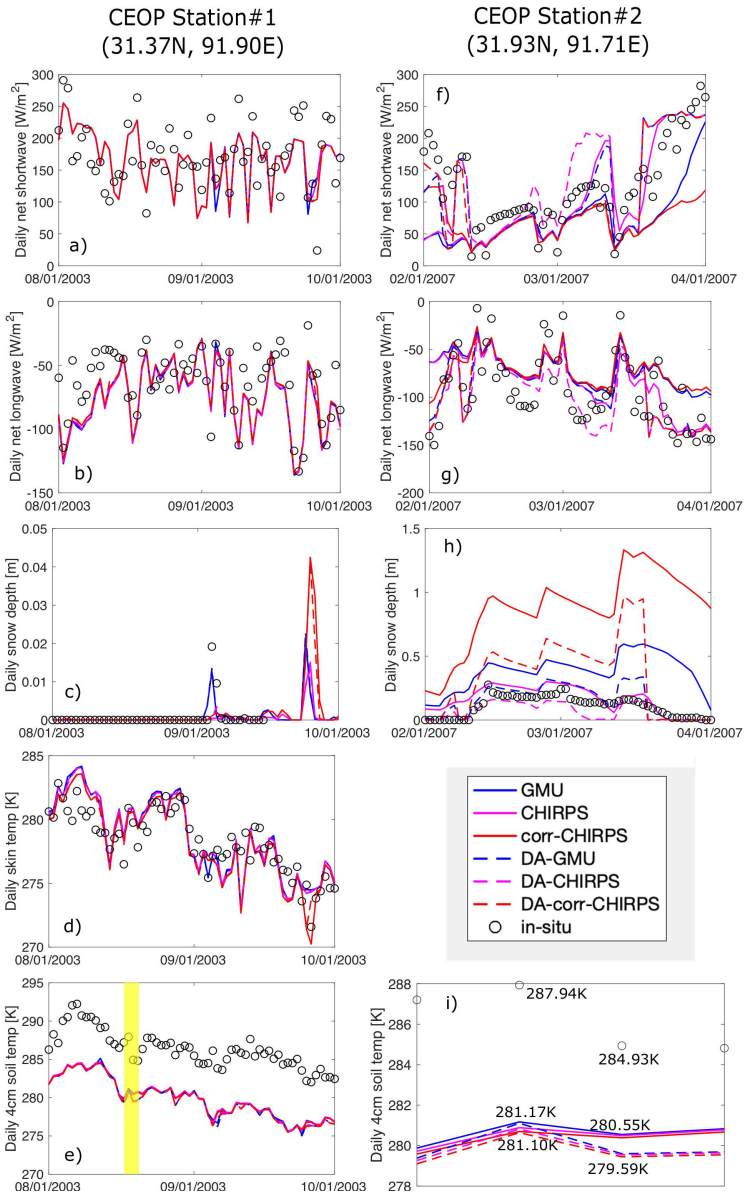

Figure 7. Example time series of all model simulations in the evaluation against two ground-based CEOP stations, a) through e) are for CEOP station#1 at 31.37°N, 91.90°E from 01 August 2003 to 01 October 2003, and f) through h) are for CEOP station#2 at 31.93°N, 91.71°E from 01 February 2007 to 01 April 2007. The 4 cm soil temperature time series from 17 August 2003 to 20 August 2003 for CEOP station#1 are highlighted and shown in i). The CEOP measurements include daily-averaged surface net shortwave radiation (row 1), surface net longwave radiation (row 2), snow depth (row 3), skin temperature (row 4), and 4 cm soil temperature (row 5). Note: no ground-based skin and soil temperature measurements were available for CEOP station#2.

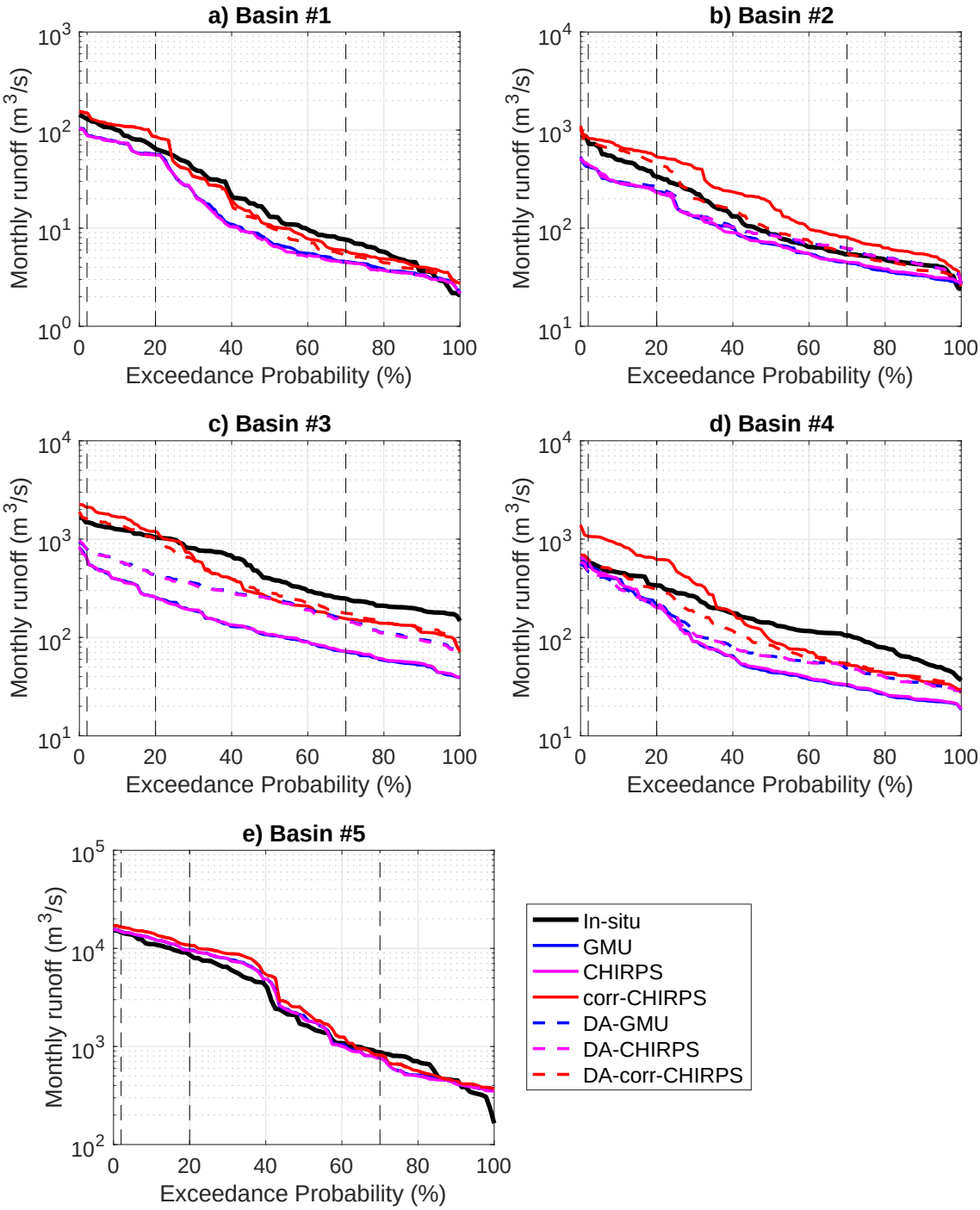

Figure 8. Flow duration curves calculated from all model simulations as well as ground-based measurements for the five gauged basins from 2003 to 2016. Note that the y-axis is plotted on a log scale. The vertical dashed lines are used to partition the curve into high-flow segment (with exceedance probability between 0% and 2%), high-to-mid flow segment (with exceedance probability between 2% and 20%), mid-flow segment (with exceedance probability between 20% and 70%), and low-flow segment (with exceedance probability between 70% and 100%).

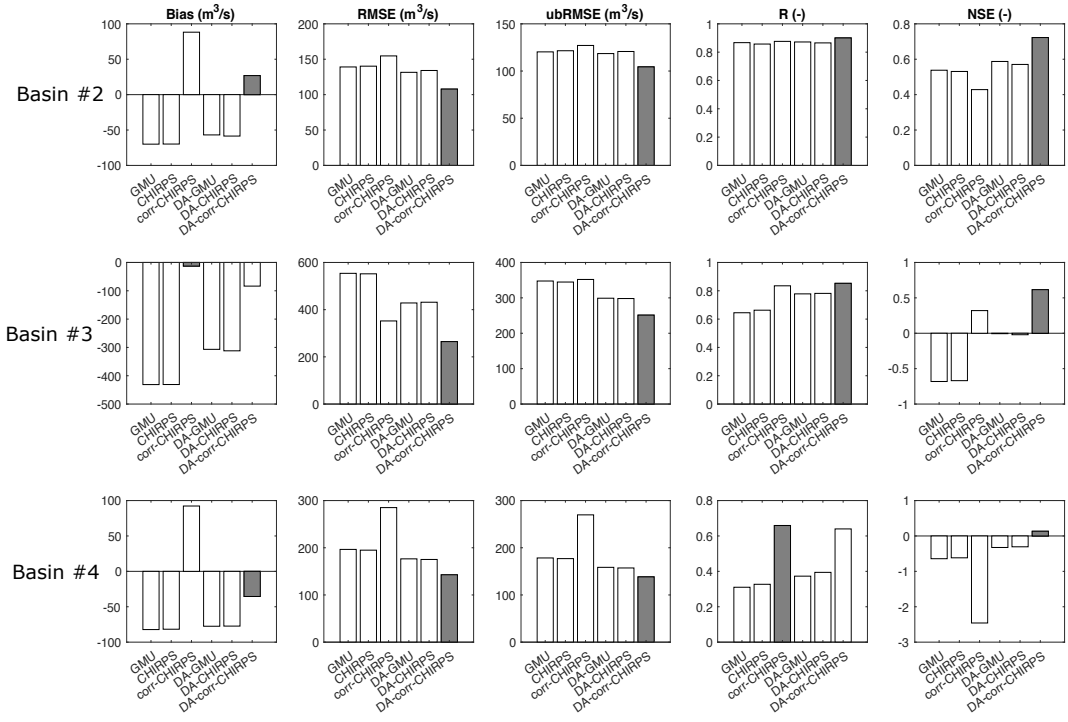

Figure 9. Statistics of bias (column 1), RMSE (column 2), ubRMSE (column 3), R (column 4), and NSE (column 5) for all experiments across three snow-dominated basins, i.e., Basin #2 (row 1), Basin #3 (row 2), and Basin #4 (row 3), in the evaluation against ground-based monthly runoff measurements from 2003 to 2016. Each row represents statistics for each basin. In addition, experiments with the best goodness-of-fit statistics for each basin are marked with grey bars.

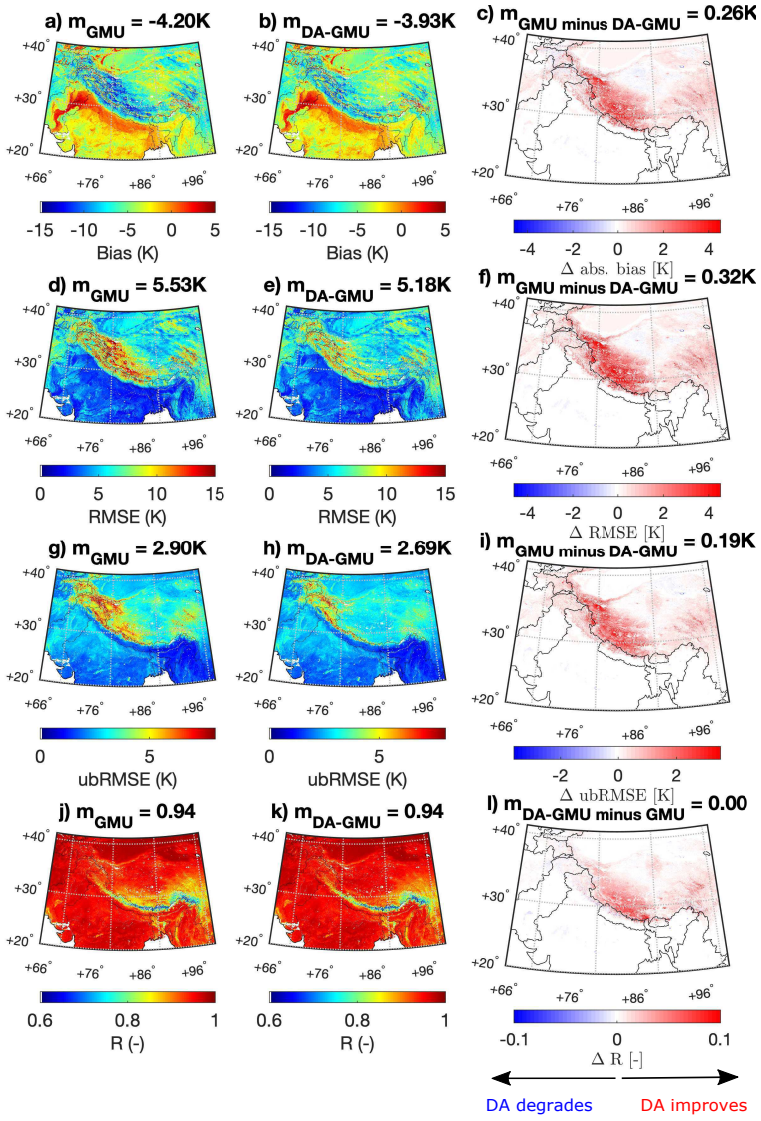

Figure 10. Spatial distribution of bias, RMSE, ubRMSE, and R computed between daily-averaged, 1-km HMA-GMU (column 1), DA-HMA-GMU (column 2) surface temperature and MODIS derived surface temperature from 2003 to 2016. Spatial distribution of the change in the absolute value of bias, RMSE, ubRMSE, and R between HMA-GMU and DA-HMA-GMU are shown in column 3. The red colors in c), f), i), and l) indicate DA-HMA-GMU derived estimates agree better with MODIS derived measurements than HMA-GMU. Conversely, blue colors indicate that HMA-GMU agrees better with MODIS. The title also demonstrates the spatial mean, m, computed for each map.

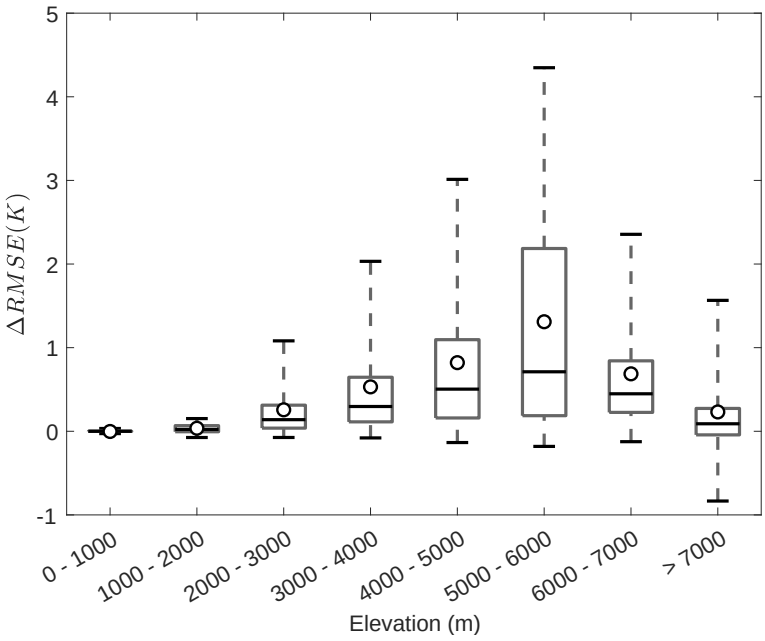

Figure 11. A box plot of ΔubRMSE computed between HMA-GMU and DA-HMA-GMU during the comparison against MODIS derived surface temperature, which are binned as a function of elevation per grid cell. The boxes show the median (marked as the black line in the box) along with the 25th and 75th percentiles while the whiskers show the 5 and 95th percentiles. The spatially-averaged skill metrics are marked as dots for each bin. The positive ΔubRMSE indicates skill improvements in the DA-HMA-GMU relative to HMA-GMU.

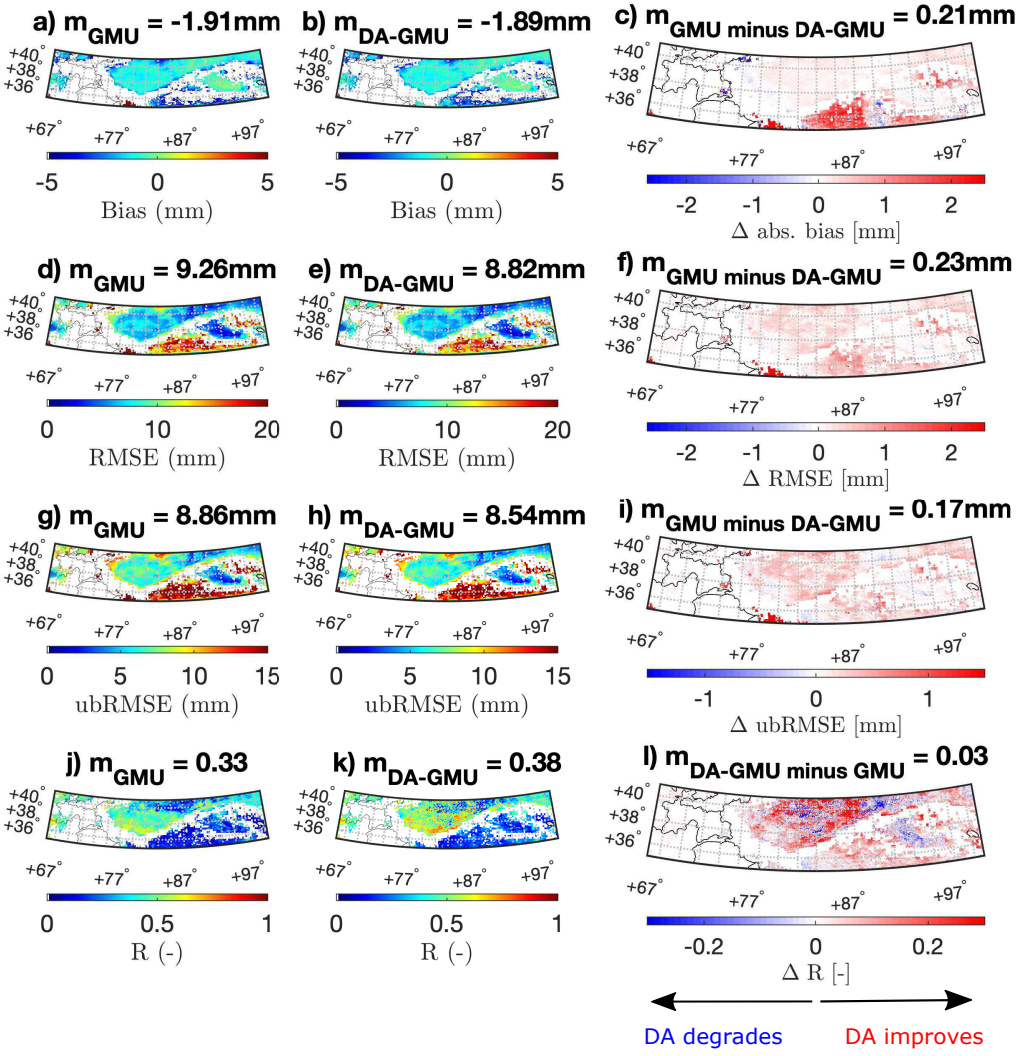

Figure 12. Same as Figure 10, but for the evaluation against daily-averaged, 5-km CGLS SWE product from 2006 to 2016. Note the domain is truncated because the CGLS SWE product only covers area above latitude 35°N, and mountainous regions are excluded from the CGLS product.

Table 1. Experiments used for evaluation. The experiment with the "DA" prefix denotes the model run with snow cover and freeze/thaw assimilatiom enabled.

Experiment name	Model output spatial resolution/temporal resolution	Precipitation input source (spatial resolution/temporal resolution)	Other meteorological forcings source (spatial resolution/temporal resolution)
HMA-GMU	0.01°/daily	Downscaled CHIRPS (0.01°/6-hourly)	Downscaled ECMWF (0.01°/6-hourly)
HMA-CHIRPS	0.01°/daily	CHIRPS (0.05°/daily)	Downscaled ECMWF (0.01°/6-hourly)
HMA-corr-CHIRPS	0.01°/daily	Bias-corrected CHIRPS (0.05°/daily)	Downscaled ECMWF (0.01°/6-hourly)
DA-HMA-GMU	0.01°/daily	Downscaled CHIRPS (0.01°/6-hourly)	Downscaled ECMWF (0.01°/6-hourly)
DA-HMA-CHIRPS	0.01°/daily	CHIRPS (0.05°/daily)	Downscaled ECMWF (0.01°/6-hourly)
DA-HMA-corr-CHIRPS	0.01°/daily	Bias-corrected CHIRPS (0.05°/daily)	Downscaled ECMWF (0.01°/6-hourly)

Table 2. The computed means and medians of each goodness-of-fit statistics metric in the evaluation against CEOP snow depth measurements across all experiments (see Figure 3 (row 4)).

Experiment name	Statistics	Mean	Median
HMA-GMU	Bias (m)	0.02	0.001
	RMSE (m)	0.05	0.01
	ubRMSE (m)	0.05	0.01
	R (-)	0.27	0.22
HMA-CHIRPS	Bias (m)	0.003	0.001
	RMSE (m)	0.03	0.01
	ubRMSE (m)	0.03	0.01
	R (-)	0.31	0.28
HMA-corr-CHIRPS	Bias (m)	0.12	0.003
	RMSE (m)	0.19	0.02
	ubRMSE (m)	0.14	0.02
	R (-)	0.24	0.17
DA-HMA-GMU	Bias (m)	0.0004	0.0006
	RMSE (m)	0.03	0.01
	ubRMSE (m)	0.03	0.01
	R (-)	0.29	0.22
DA-HMA-CHIRPS	Bias (m)	-0.003	-0.0003
	RMSE (m)	0.02	0.01
	ubRMSE (m)	0.02	0.01
	R (-)	0.31	0.25
DA-HMA-corr-CHIRPS	Bias (m)	0.006	0.002
	RMSE (m)	0.04	0.01
	ubRMSE (m)	0.04	0.01
	R (-)	0.30	0.28

Table 3. Three relative bias related statistics computed for all model simulations in the evaluation across three snow and glacier dominated gauged basins. Statistics obtained from DA-enabled experiments with greater-than-five-percent improvements relative to the non-DA counterparts are bolded.

Experiment name	Statistics	Basin #2	Basin #3	Basin #4
HMA-GMU	rbias_{FMS}	-0.08	-0.11	0.61
	rbias_{FHV}	-0.47	-0.53	0.09
	rbias_{FMM}	-0.05	-0.22	-0.23
HMA-CHIRPS	rbias_{FMS}	-0.10	-0.13	0.54
	rbias_{FHV}	-0.45	-0.53	0.10
	rbias_{FMM}	-0.04	-0.22	-0.22
HMA-corr-CHIRPS	rbias_{FMS}	0.04	0.42	1.11
	rbias_{FHV}	0.10	0.38	1.14
	rbias_{FMM}	0.16	-0.08	-0.08
DA-HMA-GMU	rbias_{FMS}	-0.21	-0.27	**0.29**
	rbias_{FHV}	**-0.44**	**-0.45**	**-0.07**
	rbias_{FMM}	**-0.00**	**-0.08**	**-0.16**
DA-HMA-CHIRPS	rbias_{FMS}	-0.24	-0.26	**0.29**
	rbias_{FHV}	-0.44	**-0.44**	**-0.01**
	rbias_{FMM}	**-0.00**	**-0.08**	**-0.16**
DA-HMA-corr-CHIRPS	rbias_{FMS}	0.16	**0.23**	**0.50**
	rbias_{FHV}	**0.03**	**0.08**	**0.18**
	rbias_{FMM}	**0.03**	**-0.06**	-0.11

Acknowledgments

This work was funded by NASA High Mountain Asia (NNH15ZDA001N-HMA) program, grant number NNX16AQ89G. Yiwen Mei is currently supported by the M-cube program (U064214) and the Water Theme project (U068408) of University of Michigan. We thank three anonymous reviewers, whose comments significantly improved the quality of the manuscript. We thank Shruti Mishra at the Argonne National Laboratory for providing runoff measurements at Basin #1 and Basin #2 obtained from the Department of Hydrology and Meteorology in Nepal. We thank Jing Tao at Lawrence Berkeley National Laboratory for sharing the codes related to shortwave downscaling. The CHIPRS precipitation data are available from `ftp://ftp.chg.ucsb.edu/pub/org/chg/products/`. The bias correction factors as applied to CHIRPS precipitation product is obtained from `http://www.gloh2o.org/pbcor/`. The CHARIS data were obtained from `http://himatmap.apps.nsidc.org/hma_insitu.html`. The GRDC data were obtained from the Global Runoff Data Centre, 56068 Koblenz, Germany (`https://www.bafg.de/GRDC/EN/01_GRDC/grdc_node.html`). The GSOD data were obtained from `https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod`. The CMA data were obtained from `https://data.cma.cn/en/?r=data/detail&dataCode=SURF_CLI_CHN_MUL_DAY_CES_V3.0&keywords=daily` as indicated in Xue et al. (2021). As of 02/14/2022, this link is no longer valid. The new link for CMA data is `http://101.200.76.197/en/?r=data/detail&dataCode=SURF_CLI_CHN_MUL_DAY_CES_V3.0`. The CEOP data were obtained from `https://www.eol.ucar.edu/projects/ceop/dm/insitu/sites/ceop_ap/`. The CTP-SMTMN data were provided by Data Assimilation and Modeling Center for Tibetan Multi-spheres, Institute of Tibetan Plateau Research, Chinese Academy of Sciences. The SETORS data were originally obtained from `http://en.tpedatabase.cn/portal/MetaDataInfo.jsp?MetaDataId=197` as indicated in Xue et al. (2021). As of 01/04/2022, this link is no longer valid. The new link for SETORS data is `https://data.tpdc.ac.cn/en/data/49ac37ac-0fc3-460f-83c4-c44744205474/`. All MODIS products were obtained from `https://earthdata.nasa.gov/`. The MEaSUREs landscape freeze/thaw product was obtained from `https://nsidc.org/data/nsidc-0728`. The CGLS SWE product (v1.0.2) was obtained from `https://land.copernicus.eu/global/products/swe`. We thank the entire NASA HiMAT team for sharing dataset and providing useful comments to the study. The NASA Land Information System was run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA

(`http://orc.gmu.edu`). The downscaling framework is implemented by functions/codes available via Mei's GitHub at `https://github.com/YiwenMei/AtmDS` and `https://github.com/YiwenMei/PrecipDS`. Downscaled products are available at NASA Distributed Active Archive Center (DAAC) at National Snow and Ice Data Center (`doi:https://doi.org/10.5067/CRN0E7YPPFGY`).

References

Arsenault, K. R., Houser, P. R., De Lannoy, G. J., & Dirmeyer, P. A. (2013). Impacts of snow cover fraction data assimilation on modeled energy and moisture budgets. *Journal of Geophysical Research: Atmospheres*, *118*(14), 7489–7504.

Beck, H. E., Wood, E. F., McVicar, T. R., Zambrano-Bigiarini, M., Alvarez-Garreton, C., Baez-Villanueva, O. M., ... Karger, D. N. (2020). Bias correction of global high-resolution precipitation climatologies using streamflow observations from 9372 catchments. *Journal of Climate*, *33*(4), 1299–1315.

Farhadi, L., Reichle, R. H., De Lannoy, G. J., & Kimball, J. S. (2015). Assimilation of freeze–thaw observations into the nasa catchment land surface model. *Journal of Hydrometeorology*, *16*(2), 730–743.

Gehne, M., Hamill, T. M., Kiladis, G. N., & Trenberth, K. E. (2016). Comparison of global precipitation estimates across a range of temporal and spatial scales. *Journal of Climate*, *29*(21), 7773–7795.

GLIMS, & NSIDC. (2005). Global land ice measurements from space glacier database (2005, updated 2018). compiled and made available by the international glims community and the national snow and ice data center. *Boulder CO, U.S.A.*. doi: 10.7265/N5V98602

Guo, Z., Dirmeyer, P. A., Hu, Z.-Z., Gao, X., & Zhao, M. (2006). Evaluation of the second global soil wetness project soil moisture simulations: 2. sensitivity to external meteorological forcing. *Journal of Geophysical Research: Atmospheres*, *111*(D22).

Hall, D., & Riggs, G. (2016). Modis/terra snow cover daily l3 global 500m grid, version 6. *Boulder, Colorado USA: National Snow and Ice Data Center*.

Hofmann, E., & Friedrichs, M. (2001). Biogeochemical data assimilation. In J. H. Steele (Ed.), *Encyclopedia of ocean sciences* (p. 302-308). Oxford: Academic Press. Retrieved from `https://www.sciencedirect.com/science/article/pii/B012227430X004104` doi: https://doi.org/10.1006/rwos.2001.0410

Kidd, C., & Huffman, G. (2011). Global precipitation measurement. *Meteorological Applications*, *18*(3), 334–353.

Kim, Y., Kimball, J., Glassy, J., & McDonald, K. (2018). Measures northern hemisphere polar ease-grid 2.0 daily 6 km land freeze/thaw status from amsr-e

and amsr2, version 1. *Boulder, Colorado USA: National Snow and Ice Data Center*.

Kim, Y., Kimball, J. S., Glassy, J., & Du, J. (2017). An extended global earth system data record on daily landscape freeze–thaw status determined from satellite passive microwave remote sensing. *Earth System Science Data*, *9*(1), 133–147.

Maggioni, V., Nikolopoulos, E. I., Anagnostou, E. N., & Borga, M. (2017). Modeling satellite precipitation errors over mountainous terrain: The influence of gauge density, seasonality, and temporal resolution. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(7), 4130–4140.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., & Gupta, H. (2015). Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, *51*(1), 716–728.

Nandakumar, N., & Mein, R. G. (1997). Uncertainty in rainfallrunoff model simulations and the implications for predicting the hydrologic effects of land-use change. *Journal of Hydrology*, *192*(1-4), 211–232.

Pulliainen, J. (2006). Mapping of snow water equivalent and snow depth in boreal and sub-arctic zones by assimilating space-borne microwave radiometer data and ground-based observations. *Remote Sensing of Environment*, *101*(2), 257–269. doi: 10.1016/j.rse.2006.01.002

Reichle, R. H., Kumar, S. V., Mahanama, S. P., Koster, R. D., & Liu, Q. (2010). Assimilation of satellite-derived skin temperature observations into land surface models. *Journal of Hydrometeorology*, *11*(5), 1103–1122.

Rodell, M., & Houser, P. (2004). Updating a land surface model with modis-derived snow cover. *Journal of Hydrometeorology*, *5*(6), 1064–1075.

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J. P., ... Bojkov, B. (2011). Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements. *Remote Sensing of Environment*, *115*(12), 3517–3529. doi: 10.1016/j.rse.2011.08.014

Wan, Z., Hook, S. J., & Hulley, G. C. (2015). Modis/terra land surface temperature/emissivity daily l3 global 1km grid, version 6. *NASA EOSDIS LP*

DAAC.

Xue, Y., Houser, P. R., Maggioni, V., Mei, Y., Kumar, S. V., & Yoon, Y. (2019). Assimilation of satellite-based snow cover and freeze/thaw observations over high mountain asia. *Frontiers in Earth Science*, *7*, 115. Retrieved from https://www.frontiersin.org/article/10.3389/feart.2019.00115 doi: 10.3389/feart.2019.00115

Xue, Y., Houser, P. R., Maggioni, V., Mei, Y., Kumar, S. V., & Yoon, Y. (2021). Evaluation of high mountain asia - land data assimilation system (version 1) from 2003 to 2016, part i: A hyper-resolution terrestrial modeling system. *Journal of Geophysical Research: Atmospheres*, *n/a*(n/a), e2020JD034131. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020JD034131 doi: https://doi.org/10.1029/2020JD034131

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the nws distributed hydrologic model. *Water Resources Research*, *44*(9).

Yilmaz, K. K., Hogue, T. S., Hsu, K.-l., Sorooshian, S., Gupta, H. V., & Wagener, T. (2005). Intercomparison of rain gauge, radar, and satellite-based precipitation estimates with emphasis on hydrologic forecasting. *Journal of Hydrometeorology*, *6*(4), 497–517.

Yoon, Y., Kumar, S. V., Forman, B. A., Zaitchik, B., Kwon, Y., Qian, Y., ... others (2019). Evaluating the uncertainty of terrestrial water budget components over high mountain asia. *Frontiers in Earth Science*, *7*, 120.

Zheng, D., Van der Velde, R., Su, B., Wen, J., & Wang, X. (2017, 02). Assessment of noah land surface model with various runoff parameterizations over a tibetan river. *Journal of Geophysical Research: Atmospheres*, *122*. doi: 10.1002/2016JD025572

Yang, K., Qin, J., Zhao, L., Chen, Y., Tang, W., Han, M., ... others (2013). A multiscale soil moisture and freeze–thaw monitoring network on the third pole. *Bulletin of the American Meteorological Society*, *94*(12), 1907–1916.