Evaluation of High Mountain Asia -Land Data Assimilation System Part I: A hyper-resolution terrestrial modeling system

Yuan Xue[1], Paul Houser[1], Viviana Maggioni[1], Yiwen Mei[2], Sujay Kumar[3], and Yeosang Yoon[3]

[1]George Mason University
[2]University of Michigan
[3]NASA/GSFC

November 21, 2022

Abstract

This first paper of the two-part series focuses on demonstrating the predictability of a hyper-resolution, offline terrestrial modeling system used for the High Mountain Asia (HMA) region. To this end, this study systematically evaluates four sets of model simulations at point scale, basin scale, and domain scale obtained from different spatial resolutions including 0.01 degree (1-km) and 0.25 degree (25-km). The assessment is conducted via comparisons against ground-based observations and satellite-derived reference products. 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, and total runoff. In the evaluation against ground-based measurements, the superiority of the 0.01 degree estimates are mostly demonstrated across relatively complex terrain. Specifically, hyper-resolution modeling improves the skill in meteorological forcing estimates (except precipitation) by 9% relative to coarse-resolution estimates. The model forced by downscaled forcings in its entirety yields the highest predictability skill in model output states as well as precipitation, which improves the skill obtained by coarse-resolution estimates by 7%. These findings, on one hand, corroborate the importance of employing the hyper-resolution versus coarse-resolution modeling in areas characterized by complex terrain. On the other hand, by evaluating four sets of model simulations forced with different precipitation products, this study emphasizes the importance of accurate hyper-resolution precipitation products to drive model simulations.

Evaluation of High Mountain Asia - Land Data Assimilation System (version 1) from 2003 to 2016, Part I: A hyper-resolution terrestrial modeling system

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 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 predictability of a hyper-resolution, offline terrestrial modeling system used for the High Mountain Asia region is presented.
- The study emphasizes the importance of using hyper-resolution versus coarse-resolution modeling in areas characterized by complex terrain.
- The study emphasizes the importance of an accurate hyper-resolution precipitation product used to drive model simulations.

Corresponding author: Yuan Xue, yxue12@terpmail.umd.edu

Abstract

This first paper of the two-part series focuses on demonstrating the predictability of a hyper-resolution, offline terrestrial modeling system used for the High Mountain Asia (HMA) region. To this end, this study systematically evaluates four sets of model simulations at point scale, basin scale, and domain scale obtained from different spatial resolutions including 0.01° ($\sim$ 1-km) and 0.25° ($\sim$ 25-km). The assessment is conducted via comparisons against ground-based observations and satellite-derived reference products. 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, and total runoff. In the evaluation against ground-based measurements, the superiority of the 0.01° estimates are mostly demonstrated across relatively complex terrain. Specifically, hyper-resolution modeling improves the skill in meteorological forcing estimates (except precipitation) by 9% relative to coarse-resolution estimates. The model forced by downscaled forcings in its entirety yields the highest predictability skill in model output states as well as precipitation, which improves the skill obtained by coarse-resolution estimates by 7%. These findings, on one hand, corroborate the importance of employing the hyper-resolution versus coarse-resolution modeling in areas characterized by complex terrain. On the other hand, by evaluating four sets of model simulations forced with different precipitation products, this study emphasizes the importance of accurate hyper-resolution precipitation products to drive model simulations.

1 Introduction

High Mountain Asia (HMA) forms the headwaters of river systems, e.g., Yangtze, Yellow, Mekong, Brahmaputra, Indus, and Ganges Rivers, that provide fresh water supply for more than a billion people in the region for the purposes of downstream irrigation, hydropower generation, and general consumption (Armstrong et al., 2019). Meteorological and hydrological conditions in such mountainous environment are poorly monitored due to terrain inaccessibility and financial insufficiency (Ghatak et al., 2018). To overcome the limitations imposed by inadequate ground-based stations, previous studies generally utilized global land surface models or regional hydrological models to represent the hydro-meteorological processes involved across the HMA region. For example, Immerzeel, Droogers, De Jong, and Bierkens (2009) evaluated runoff simulations in a Himalayan river basin using the Snowmelt Runoff Model forced by remotely sensed precipitation at a spatial resolution of 0.25°. Yoon et al. (2019) provided a thorough evaluation of the terrestrial water budget estimation (i.e., precipitation, evapotranspiration, runoff, and terrestrial water storage) over HMA using a suite of uncoupled global land surface models at a spatial resolution of 0.25°. Further, the study conducted by Ghatak et al. (2018) evaluated the Noah land surface model-derived runoff simulations in a HMA region at a spatial resolution of 5-km. To our current knowledge, there exists no published study performing land surface model simulations finer than 5-km for the entire HMA for a relatively long period (e.g., more than 10 years).

As pointed out by Singh, Reager, Miller, and Famiglietti (2015), increasing computational efficiency and the need for improved accuracy are driving the development of "hyper-resolution" land surface models that can be implemented at regional scales, with spatial resolutions of 1-km or even finer. In addition, previous studies emphasized that high spatial heterogeneity over complex terrain requires land surface model simulations to be implemented at relatively high spatial resolutions (e.g., Zhao and Li (2015)). In addition to the tremendous amount of computational resources, one of the primary challenges of land surface modeling at hyper-resolution is the lack of forcing datasets at such resolution (Kollet et al., 2010; Singh et al., 2015). That is, we simply do not have reliable regional-scale 1-km in-situ or satellite observational capabilities from which to derive all meteorological forcing variables required as input into land surface models. Thanks to the recent developments in physical, and statistical downscaling approaches (e.g., Mei, Maggioni, Houser, Xue, and Rouf (2020); Rouf, Mei, Maggioni, Houser, and Noonan (2019)), which allows hyper-resolution forcing fields to be derived from coarser-resolution data based on ancillary information (e.g., land cover, surface roughness, and topography). Using Yoon et al. (2019) as a benchmark, in this study, we attempt to address the following science question: "to what extent does the development of hyper-resolution forcing input improve or worsen land surface modeling, compared to ground-based observations or satellite-derived reference products"? To this end, this study systematically evaluates the 0.01° ($\sim$ 1-km) and 0.25° ($\sim$ 25-km) model simulations at point-scale, basin-scale, and domain-scale. The key variables of interest include various downscaled meteorological forcing input, as well as model output of surface net shortwave radiation, surface net longwave radiation, skin temperature, near-surface soil temperature, snow depth, snow water equivalent, and total runoff.

The ultimate goal of this research is to evaluate the newly-developed, hyper-resolution High Mountain Asia - Land Data Assimilation System (version 1) from 2003 to 2016. The High Mountain Asia - Land Data Assimilation System is intended to provide spatially and temporally continuous land surface estimates, which are believed essential to capture the spatio-temporal evolution of hydrometeorological conditions and their associated processes across HMA. Part I, presented in this manuscript, focuses on demonstrating the predictability of a hyper-resolution (at $\sim$ 1-km spatial resolution), offline (uncoupled to the atmosphere) terrestrial modeling system (without assimilation) used for complex terrain regions.

2 Data and Methods

2.1 Study domain and models

The study domain is the HMA region bounded between 20°N and 41°N and 66°E and 101°E. Meteorological fields from the European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni, Buizza, Palmer, and Petroliagis (1996)) and Climate Hazards Group InfraRed Precipitation with Station data, Version 2 (CHIRPS; Funk et al. (2015)) (and two precipitation variants derived from CHIRPS; see Table 1) are used in this study. The ECMWF product is originally on a TL511 triangular truncation, linear reduced gaussian grid (0.25°) for four synoptic hours: 00, 06, 12, and 18 UTC. The ECMWF forcing fields employed in this study include air temperature, specific humidity, downward longwave flux, downward shortwave flux, wind speed, and surface pressure. The CHIRPS precipitation product has a native spatial resolution of 0.05° at a daily time scale. Yoon et al. (2019) demonstrated that the joint use of ECMWF and CHIRPS forcings provides the best model estimates at 0.25° spatial resolution for daily output of water balance components.

Four sets of model simulations are evaluated in this study, which are summarized in Table 1. 1) In "HMA-Coarse" (also denoted as "HMA-CS" in figures), the meteorological inputs (i.e., air temperature, humidity, surface pressure, wind, downward shortwave, and longwave radiation) are adjusted for the elevation differences through lapse-rate and slope-aspect correction methods (Kumar, Peters-Lidard, Mocko, & Tian, 2013). Inputs obtained from ECMWF and CHIRPS are spatially interpolated and aggregated onto the same 0.25° grid for generating model output. 2) In "HMA-GMU", all meteorological inputs are downscaled using physically-based and statistically-based algorithms onto the 0.01° grid for model estimates. Section 2.1.1 summarizes key steps used in the downscaling process. 3) In "HMA-CHIRPS", model output estimates are at a spatial resolution of 0.01°. Except for the precipitation field, all other meteorological forcings remain the same as "HMA-GMU". The precipitation field is replaced with original CHIRPS, which is then spatially interpolated onto the same 0.01° grid for model estimates using the simplistic conservative interpolation scheme. 4) In "HMA-corr-CHIRPS", model output estimates are at a spatial resolution of 0.01°. Except for precipitation, all other meteorological forcings remain the same as "HMA-GMU" and "HMA-CHIRPS". The precipitation field is replaced with the bias-corrected CHIRPS (see Section 2.1.2 for details), which is then spatially interpolated onto the same 0.01° grid for model estimates using the simplistic conservative interpolation scheme.

The land surface model used in this study is the baseline Noah-MP (Niu et al., 2011; Z.-L. Yang et al., 2011). Noah-MP is enhanced from the original Noah land surface model through the addition of improved model physics (i.e., dynamic vegetation phenology, a carbon budget and carbon-based photosynthesis, an explicit vegetation canopy layer, a multilayer snowpack representation and a groundwater module) and multi-parameterization options. We used Noah-MP version 3.6 within the NASA Land Information System (LIS) 7.2 version (Kumar et al., 2006). The Noah-MP model configuration options are the same as Xue et al. (2019), and Yoon et al. (2019), which were shown to provide relatively good agreement with reference datasets in simulating hydrological conditions. The land surface model simulations are conducted with a 15-min time step for a 14-year time period (2003–2016) to generate daily output of water balance components. The initial conditions for the runs are generated by appropriate spin-up strategies as described by Xue et al. (2019) and Yoon et al. (2019), and then reinitializing all model runs in 2003.

2.1.1 Downscaling of meteorological forcings

Following Rouf et al. (2019), meteorological forcings including near-surface ($\sim$ 10 m above the ground) air temperature (denoted as "T_a"), surface pressure (denoted as "p_r"), near-surface ($\sim$ 10 m above the ground) specific humidity (denoted as "q"), near-surface ($\sim$ 10 m above the ground) wind speed (denoted as "w"), downward surface shortwave radiation (denoted as "SW"), and downward surface longwave radiation (denoted as "LW") obtained from ECMWF are spatially downscaled from their original resolutions (0.25°) onto the 0.01° model grid. The symbol of "$\tilde{(\cdot)}$" denotes the variable at 0.01° model grid. The downscaling methods are developed by the George Mason University (GMU) research team, and therefore we refer to the downscaled meteorological forcings as GMU downscaled forcings. The downscaled air temperature in the unit of K is computed as Marshall and Plumb (1989):

$$\tilde{T_a} = T_a + \Gamma_a(\tilde{Z} - Z), \tag{1}$$

where Z (m) is the Shuttle Radar Topography Mission (SRTM) digital elevation model derived elevation at 0.25°, $\tilde{Z}$ (m) is the elevation derived at 0.01° (see Figure 1a), and Γ_a (K/m) is the spatially distributed dynamic lapse rate in air temperature (Rouf et al., 2019). The downscaled surface pressure in the unit of Pa is computed as Cosgrove et al. (2003):

$$\tilde{p_r} = p_r \exp(-\frac{g(\tilde{Z} - Z)}{RT_m}), \tag{2}$$

where exp(·) is the exponential operator. R (= 287 J/(kg · K)) is the ideal gas constant, g (= 9.81 m/s^2) is the gravitational acceleration constant, and T_m (K) is the mean air temperature computed from T_a and $\tilde{T_a}$. The downscaled specific humidity in the unit of kg/kg is computed as Lawrence (2005):

$$\tilde{q} = \frac{0.622\tilde{E}}{\tilde{p_r} - 0.378\tilde{E}}, \tag{3}$$

where

$$\tilde{E} = C_1 \exp \frac{C_2\tilde{T_d}}{\tilde{T_d} + C_3}, \tag{4}$$

$$\tilde{T_d} = T_d + \Gamma_d(\tilde{Z} - Z), \tag{5}$$

where for water, C_1 (= 611.21 Pa), C_2 (= 17.268), C_3 (= 238.88°C), and for ice, C_1 (= 611.15 Pa), C_2 (= 22.452), C_3 (= 272.55°C) as noted in Buck (1981). T_d (K) is the dew point temperature, and Γ_d (K/m) is the spatially distributed dynamic lapse rate in dew point temperature. The downscaled wind speed in the unit of m/s is computed as Bohn and Vivoni (2019); Rouf et al. (2019); Tao and Barros (2018):

$$\tilde{w} = \frac{\tilde{\mu_*}}{\kappa} \ln \frac{H}{\tilde{z_0}}, \tag{6}$$

where

$$\tilde{\mu_*} = \mu_*(\frac{\tilde{z_0}}{z_0})^{0.09}, \tag{7}$$

$$\tilde{z_0} = \tilde{k}\sum_{i=1}^{M}\tilde{\rho}_i z_{0,i} + z_0 - k\sum_{i=1}^{M}\rho_i z_{0,i}, \tag{8}$$

where ln(·) is the natural logarithm operator, μ_* (m/s) is the friction velocity, z_0 (m) is the surface roughness, κ (= 0.41) is the Von Kármán constant, H (= 10 m) is the measurement height above the ground, and M is the number of land cover types. ρ_i is the fractional values of the ith land cover type. k represents the temporal variability of the Moderate Resolution Imaging Spectroradiometer (MODIS) derived normalized difference vegetation index (NDVI), which is computed as the ratio of the NDVI obtained from the current time step versus the annual mean of the NDVI. The downscaled incident shortwave radiation in the unit of W/m^2 is computed as Fiddes and Gruber (2014); Gupta and Tarboton (2016); Ruiz-Arias, Alsamamra, Tovar-Pescador, and Pozo-Vázquez (2010); Tao and Barros (2018):

$$\tilde{SW} = \delta \cos(\theta) \exp(\tau(\tilde{p_r} - p_r))SW_b + F_v SW_d + \alpha F_t(\tilde{SW_b} + (1 - F_v)\tilde{SW_d}), \tag{9}$$

where SW_b (W/m^2) is the direct shortwave radiation, and SW_d (W/m^2) is the diffuse shortwave radiation. δ is the binary shadowing mask indicating whether the grid cell is blocked by the shadow of nearby terrain, $\cos(\theta)$ is the cosine of the solar illumination angle, τ (Pa^{-1}) is the broadband attenuation coefficient, α is the MODIS derived surface albedo, F_v is the fractional value of the visible sky, and F_t is the terrain configuration factor, which is computed as the function of terrain slope and F_v. The downscaled longwave radiation in the unit of W/m^2 is computed as Fiddes and Gruber (2014); Konzelmann et al. (1994):

$$\tilde{LW} = (\tilde{\epsilon_c} + \Delta\epsilon)\sigma \tilde{T_a}^4, \tag{10}$$

where

$$\tilde{\epsilon_c} = 0.23 + 0.484(\frac{\tilde{E}}{\tilde{T_a}})^{\frac{1}{8}}, \tag{11}$$

$$\Delta\epsilon = \frac{LW}{\sigma {T_a}^4} - \epsilon_c, \tag{12}$$

where σ (= 5.67 × 10^{-8} W/(m^2 · K^4)) is the Stefan-Boltzmann constant, and ϵ_c is the clear-sky emissivity.

The original 0.05°/daily CHIRPS precipitation is spatially and temporally downscaled to 0.01°/6-hourly by weighting factors. To disaggregate CHIRPS to 0.01°, spatially-distributed weighting factors are derived from daily cumulative downscaled 0.01° ECMWF precipitation, which is derived from the original 0.25°/6-hourly ECMWF precipitation following Mei et al. (2020). The kernel of the Mei et al. (2020) precipitation downscaling framework lies in a random forest (RF) classification along with a regression algorithm. The framework first applies the recursive feature elimination algorithm to select important predictors in terms of their predictive values to the daily cumulative ECMWF precipitation from a list of potential predictors. There are 13 potential predictors including eight meteorological variables (air and dew point temperature, surface pressure, specific and relative humidity, longwave and shortwave radiation, and wind speed) and five auxiliary variables (vegetation index with 30-day and 60-day lag, latitude, longitude, and day of year). The meteorological variables are either adopted or derived from the downscaled 0.01° ECMWF estimates. For each year from 2003 to 2016, the first seven predictors with higher predictive values are selected as important predictors. In a next step, with the identified predictors, RF classification models are trained to a binary precipitation mask defining rainy (i.e., daily cumulative precipitation being greater than 0 mm) and non-rainy grid cells and RF regression models are trained to the daily cumulative precipitation for rainy grid cells (Note: one RF classification and one RF regression model for a year). Then, the trained RF classification models are used to produce the 0.01° daily binary precipitation masks with the 0.01°/daily predictors. Finally, the RF regression models are used to estimate the daily cumulative precipitation for rainy grid cells (inferred by the 0.01° precipitation masks) with the identified predictors.

After attaining the 0.01°/daily ECMWF precipitation, the 0.05°/daily CHIRPS precipitation is spatially disaggregated following the equations below:

$$\tilde{pd}_C = \begin{cases} \frac{\tilde{pd}_{E,i}}{\frac{1}{N}\sum_{i=1}^{N}\tilde{pd}_{E,i}} pd_C, & if\, \frac{1}{N}\sum_{i=1}^{N}\tilde{pd}_{E,i} > 0, \\ pd_C, & if\, \frac{1}{N}\sum_{i=1}^{N}\tilde{pd}_{E,i} = 0 \end{cases} \tag{13}$$

where pd_E and pd_C represent the daily cumulative precipitation from ECMWF and CHIRPS, respectively. N is the total number of 0.01° grid cells within a 0.05° grid cell. The term of $\frac{\tilde{pd}_{E,i}}{\frac{1}{N}\sum_{i=1}^{N}\tilde{pd}_{E,i}}$ denotes the spatially distributed weighting factors, which quantifies the 0.01° variability of precipitation within the 0.05° grid cells. In the case that all 0.01° grid cells within a 0.05° grid cell have null precipitation, pd_C is distributed evenly. The daily cumulative CHIRPS precipitation is then multiplied by a temporal weighting factor to attain the 6-hourly precipitation value at 0.01° (denoted as "$\tilde{pt}_C$"). The temporal weighting factor is derived from the 0.25°/6-hourly ECMWF precipitation, written as:

$$\tilde{pt}_C = \begin{cases} \frac{pt_{E,t}}{\sum_{t=1}^{T} pt_{E,t}} \tilde{pd}_C, & if\, \sum_{t=1}^{T} pt_{E,t} > 0, \\ \tilde{pd}_C, & if\, \sum_{t=1}^{T} pt_{E,t} = 0 \end{cases} \tag{14}$$

where pt_E denotes the 6-hourly ECMWF precipitation. T is the total number of time steps within one day. Similar to Equation 13, the term of $\frac{pt_{E,t}}{\sum_{t=1}^{T} pt_{E,t}}$ is the 6-hourly temporal weighting factor used to distribute the daily cumulative precipitation; if all 6-hourly precipitation values are zeros within a day, $\tilde{pd}_C$ is distributed evenly.

2.1.2 Bias-corrected CHIRPS

The bias-corrected CHIRPS are generated using the original CHIRPS at 0.05° multiplied with the monthly, spatially-distributed correction factors given by Beck et al. (2020). Their study used streamflow observations from 9372 stations for calibrations of several state-of-the-art (quasi-) global precipitation climatologies. Monthly climatological bias correction factors were calculated by disaggregating the long-term bias correction factors on the basis of gauge catch efficiencies. An example of the spatially-distributed precipitation correction factors as applied in CHIRPS product in February across HMA can be seen from Figure 1b. The domain-averaged precipitation correction factor is 1.43, with relatively high correction factors presence along Karakoram and Himalayan ranges. As noted in Beck et al. (2020), these regions exhibit marked elevation gradients, sparse gauge networks, and substantial snowfall: all factors that tend to favor precipitation underestimation, and therefore, the newly-generated bias-corrected CHIRPS product is intended to increase the magnitude of precipitation across HMA (see Figure 11).

2.2 Ground-based measurements of meteorological conditions

A summary of ground-based measurements of meteorological conditions used for evaluation is listed in Tables 3 and 4. These measurements include air temperature, wind speed, specific humidity, surface pressure, incident shortwave radiation, incident longwave radiation, and total precipitation. These dataset are obtained from 1) the Chinese Meteorological Administration (CMA), namely the Dataset of Daily Climate Data From Chinese Surface Stations for Global Exchange (V3.0) (`https://data.cma.cn/en/?r=data/detail&dataCode=SURF_CLI_CHN_MUL_DAY_CES_V3.0&keywords=daily`), or 2) the Coordinated Enhanced Observing Period (CEOP) Asia Monsoon project (`https://www.eol.ucar.edu/projects/ceop/dm/insitu/sites/ceop_ap/`), or 3) the Department of Hydrology and Meteorology in Nepal (DHM), or 4) the Pakistan Meteorology Department (PMD), or 5) the weather underground (WU; `https://www.wunderground.com`). Locations of the ground-based stations are shown in Figures 3 through 5. The discrepancies between model estimates and measurements resulting from different measurement heights are neglected in this study (Note: some in-situ data source do not provide the measurement height information).

2.3 Ground-based measurements of modeled states

A summary of ground-based measurements of modeled states used for evaluation is listed in Table 4.

2.3.1 Surface radiation

Surface net shortwave radiation and net longwave radiation, calculated as incoming-minus-outgoing radiant energy fluxes, are evaluated in this study, respectively. The in-situ radiation measurements are obtained from CEOP. Radiation fluxes are measured using CM21 Kipp & Zonen (or 2770 Aandera) sensors at a time step of an hour (or twenty minutes), and at a height of 1.58 m, 2 m (or 3.1 m) above from the ground surface (depending on the station). Daily-averaged, in-situ fluxes are then computed as the temporal mean of the values collected during the 24-hour period. The measurement discrepancies as a result of different sensor installation heights are neglected in this study.

2.3.2 Skin temperature

Two different sources of skin temperature measurements are obtained. First, in-situ, daily-averaged surface temperature measurements are obtained from CMA. The daily-averaged surface temperature values are computed by averaging the four measurements taken by platinum resistance thermometers at 02:00, 08:00, 14:00, and 20:00. Second, the in-situ surface temperature measurements are obtained from the CEOP Asia Monsoon project. Skin temperature are measured at a time step of an hour. Daily-averaged, in-situ temperatures are then computed as the temporal mean of the values collected during the 24-hour period.

2.3.3 Snow depth

The in-situ, daily-averaged snow depth measurements are obtained from 1) the Global Summary of the Day (GSOD; `https://data.noaa.gov/dataset/dataset/global-surface-summary-of-the-day-gsod`), 2) the Contribution to High Asia Runoff from Ice and Snow

(CHARIS) project (http://himatmap.apps.nsidc.org/hma_insitu.html), and 3) the CEOP Asia Monsoon project.

2.3.4 Near-surface soil temperature

Three different sources of the near-surface (5 cm below the ground) soil temperature measurements are obtained. First, in-situ soil temperature measurements are obtained from the CEOP Asia Monsoon project. Near surface soil temperatures are measured at a time step of an hour or twenty minutes, and at the depth of 3 cm, 4 cm, and/or 5 cm from the ground surface (depending on the station). Daily-averaged temperature values are then computed as the temporal mean of the temperatures collected during the 24-hour period as a function of the measured depth. It is assumed that measurements taken at the depth of 5 cm (i.e., center of the soil layer) can best represent the modeled top-layer of soil (0 - 10 cm). Therefore, the Inverse Distance Weighting method is applied to the model estimates to match with the measurement depths of 3 cm and 4 cm, respectively.

Second, daily-averaged near-surface soil temperature measurements from one station located at (29.76°N, 94.74°E) are obtained from the Southeastern Tibet Observation and Research Station for the Alpine Environment (SETORS; http://en.tpedatabase.cn/portal/MetaDataInfo.jsp?MetaDataId=197) maintained by the Chinese Academy of Sciences. At this station, soil temperature at a depth of 4 cm below the ground are measured using a Campbell 107 sensor. We then interpolate the modeled top-layer of soil (0 - 10 cm) temperature estimates to 4 cm using Inverse Distance Weighting to match with the measurement depth.

Third, in-situ, daily- and spatially-averaged near-surface soil temperature measurements are obtained from the Central Tibetan Plateau Soil Moisture and Temperature Monitoring Network (CTP-SMTMN; http://dam.itpcas.ac.cn/rs/?q=data) maintained by the Institute of Tibetan Plateau Research, Chinese Academy of Science. Near-surface soil temperature measurements are taken at the soil depth in between 0 and 5 cm. Only the range of the near-surface measurement depth is given in the CTP-SMTMN document without the exact measurement depth (K. Yang et al., 2013). Therefore, the modeled top-layer soil temperature is used to approximate the measurement taken at in-situ sites.

2.3.5 Total runoff

Table 2 summarizes the main characteristics of the five gauged basins (see Figure 2) in the study area, including drainage area, data source, and mean elevation computed via averaging all grid cells coincident within the given basin. These ground-based measurements are obtained from 1) the Contribution to High Asia Runoff from Ice and Snow (CHARIS) project, or 2) Department of Hydrology and Meteorology in Nepal, or 3) the Global Runoff Data Centre, 56068 Koblenz, Germany (`https://www.bafg.de/GRDC/EN/01_GRDC/grdc_node.html`). Basin #1 through Basin #5 are listed and organized by drainage area in ascending order in Table 2. It is important to note that only basins with drainage areas of greater than 625 km^2 are included in this study.

Basin #1 originates in the higher mountains in Nepal, where monsoon precipitation constitutes the major source of discharge water. In this basin, there exists a fairly clear rainfall-runoff relationship. That is, strong commonality with precipitation highs to lows matching up with flow magnitudes tends to occur frequently (Hannah, Kansakar, Gerrard, & Rees, 2005). According to Hannah et al. (2005), the flow regime shape in Basin #1 is Class C with marked August peak runoff. The flow regime magnitude in Basin #1 is Class 2 with intermediate amount of both annual total precipitation and total runoff. Note names of "Class C" and "Class 2" are classification schemes based on Hannah et al. (2005).

Basin #2 is a trans-boundary basin lying north-south in the central Himalayan region. It extends from China in the north, and flows through Nepal. The majority of the glaciated region in Basin #2 are located in Tibet, China. The climate is dominated by the Indian summer monsoon system, with the majority of the precipitation falls between June and September. Total runoff varies throughout the year influenced by both snow (and glacier) melt and precipitation (Dandekhya et al., 2017). Peak flows generally occur in July or August as the peak snow and glacier melt coincide with the monsoon peak (Mishra et al., 2018).

Basin #3 originates in Tajikistan and flows towards Uzbekistan. The highest precipitation is often brought by Westerlies during winter and spring periods, with minimums during summer and early autumn periods (Gafurov et al., 2015). The discharge regime is strongly dominated by snow (and glacier) melt in the area during summer time. The increase of water discharge typically begins in April and peaks around July or August. The recession of the discharge river flow generally commences in August and continues until February or March, when it reaches its minimum discharge point (Kulmatov, Opp, Groll, & Kulmatova, 2013).

Basin #4 is located in Tajikistan, which is mainly fed by melting snow and glaciers. The region is under the continental climate, characterized by a wide temperature variation throughout the year, with the coldest temperature generally occurring in January. Similar to Basin #3, Mid Latitude Westerlies are the dominant climatic influence in the area. Precipitation decreases from west to east. The majority of the annual precipitation falls between February and May (Grin, Schaller, & Ehlers, 2018), while during the summer and early autumn seasons precipitation presents a minimum.

Basin #5 is located in the North Western part of Myanmar. It is dominated by a mountainous forested terrain, except for the wide flood plain at its lowest southern part (Yuan et al., 2017). Rainfall is the major driver for the discharge regime in the area. During the southwest monsoon season, Basin #5 is prone to severe floods, due to the high precipitation intensities with significant spatial and temporal variations (Yuan et al., 2017). Riverine floods are very common in Basin #5, and they occur as a result of the intense precipitation when the monsoon troughs or low pressure waves superimpose on the general monsoon pattern (Latt, 2015).

2.4 Reference remotely sensed products

A summary of remotely sensed products used for evaluation is listed in Table 5.

2.4.1 Skin temperature

Similar to the evaluation strategy described in Xue et al. (2019), 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)). Given the availability of both nighttime and daytime land surface maps generated by MOD11A1 and MYD11A1 from 2003 to 2016, we use the simple arithmetic mean of all four measurements to approximate daily-averaged values. It is important to note that when daytime MOD11A1, nighttime MOD11A1 as well as daytime MYD11A1, and nighttime MYD11A1 present simultaneously, we calculate the daily-averaged surface temperature value; otherwise, a "no-value" flag is applied.

2.4.2 Snow water equivalent

The reference satellite-based snow water equivalent (SWE) 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. The CGLS SWE retrieval algorithm combines information from satellite-based microwave radiometer and optical spectrometer observations with ground based weather station snow depth measurements and produces daily Northern Hemispherical scale SWE estimates. The SWE product covers all land surface areas between latitudes 35°N and 85°N with the exception of mountainous regions, and glaciers. Therefore, the CGLS SWE product only covers about 16.3% of the entire HMA land area.

2.5 Evaluation methods

All four 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 overlapping period from 01 February 2003 to 30 November 2016 are used for evaluation in this study. It is important to note that stations (or grid cells) with records less than 200 days are excluded from the evaluation. Evaluations are conducted at three different spatial scales. The point-scale evaluations are performed via comparisons against the closest colocated ground-based stations. That is, the performance of air temperature, wind speed, specific humidity, surface pressure, incident shortwave radiation, incident longwave radiation, total precipitation, surface 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 ground-based stations. Goodness-of-fit statistics (see Section 2.5.1) are computed and a scoring system (see Appendix A) is designed to rank the performance of different sets of estimates. It is always difficult to compare 1-km scale estimates against in-situ scale stations due to the stations' representativeness issue. Therefore, if the relative elevation difference between the 1-km scale grid cell and colocated station is greater than 50%, we deem that the station is unrepresentative of the large-scale model estimates, and thus such stations are removed from the evaluation.

The basin-scale evaluations are conducted for modeled runoff through comparisons against ground-based discharge measurements. That is, this study aggregates daily-averaged total runoff output onto monthly averages and then evaluates against ground-based discharge measurements taken at basin outlets. The main reason for comparing runoff at monthly scale, rather than at hourly and daily scales is that no river routing routines are employed in this study. For each of the model simulation listed in Table 1, the modeled basin-scale total runoff is computed by integrating the runoff output at each grid cell across each of the drainage basin. The goodness-of-fit statistics plus the Nash–Sutcliffe model efficiency coefficient (see Section 2.5.1) are computed to evaluate the modeled runoff performance.

The domain-scale evaluations are conducted between 1) model estimates and reference satellite-based products, as well as between 2) meteorological forcings before and after being downscaled. 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. All model output and reference products are aggregated onto the same 0.25° grid for this set of evaluation. All SWE estimates in June, July, and August are excluded from evaluation due to minimized coverage of snow in summertime. In addition, the performance of the downscaled meteorological forcings are evaluated using the normalized mutual information index (Section 2.5.2), which is intended to serve as a proxy for the spatial similarity between the multi-year averaged forcing variable before and after being downscaled.

2.5.1 Evaluation statistics

Goodness-of-fit statistics used for evaluation include bias, root mean squared error (RMSE), unbiased root mean squared error (ubRMSE), and correlation coefficient (R). The symbol, x_{model}, is used to denote estimates obtained from the given model simulation. The symbol, x_{meas}, is used to denote in-situ measurements (or reference satellite-based measurements) at either daily or monthly time steps (Note: monthly time step is only applicable for runoff assessment). The bias is computed as:

$$Bias = \frac{1}{N_t}\sum_{j=1}^{N_t}(x_{model,j} - x_{meas,j}), \tag{15}$$

where N_t denotes the total sample size. A lower absolute value of bias is deemed better at decreasing the systematic errors. RMSE is computed as:

$$RMSE = \sqrt{\frac{1}{N_t}\sum_{j=1}^{N_t}(x_{model,j} - x_{meas,j})^2}. \tag{16}$$

A lower RMSE reflects decreased systematic errors and random errors. Further, ubRMSE is calculated as:

$$ubRMSE = \sqrt{(RMSE)^2 - (Bias)^2}. \tag{17}$$

A lower ubRMSE reflects reduced amount of random errors. In addition, R is computed as:

$$R = \frac{\sum_{j=1}^{N_t}(x_{model,j} - \bar{x}_{model})(x_{meas,j} - \bar{x}_{meas})}{\sqrt{\sum_{j=1}^{N_t}(x_{model,j} - \bar{x}_{model})^2}\sqrt{\sum_{j=1}^{N_t}(x_{meas,j} - \bar{x}_{meas})^2}}, \tag{18}$$

where $\bar{x}_{meas}$ is the time-averaged estimates of the measurements, and $\bar{x}_{model}$ is the time-averaged estimates obtained from model simulations. A higher R demonstrates better correlations with the reference. Overall, a relatively low absolute value of bias, or low RMSE, or low ubRMSE, or high R is deemed as a higher level of accuracy in the model estimates.

In addition, we compute the Nash–Sutcliffe model efficiency coefficient (NSE) statistics (Nash & Sutcliffe, 1970) in the basin-scale runoff evaluation. NSEs are used to emphasize peak values in evaluating simulation fit, which can be a useful indicator to distinguish the skills among different experiments for peak runoff predictability. NSEs can range from -infinity to 1.0. An NSE of 1.0 corresponds to a perfect match between model and observed runoff, whereas an NSE less than 0 occurs when the model simulations are not better than solely the mean of the observations.

2.5.2 Spatial similarity assessments for downscaled products

Mutual information – without an upper bound – can be used to quantify the statistical information shared between two distributions (Cover & Thomas, 1991; Strehl & Ghosh, 2002), provides a sound indication of the shared information between two dataset. On top of that, the normalized mutual information (NMI) could be further derived as a proxy for spatial similarity, which is the normalization of the mutual information index to scale the results between 0 (no correlation) and 1 (perfect correlation). That is, the NMI close to zero indicates high dissimilarity between the two distributions, whereas the NMI close to one indicates high similarity.

Following Strehl and Ghosh (2002), we define the NMI between variable $\mathbf{X}$ and $\mathbf{Y}$ as follows:

$$NMI(\mathbf{X}, \mathbf{Y}) = \frac{I(\mathbf{X}; \mathbf{Y})}{\sqrt{H(\mathbf{X})H(\mathbf{Y})}}, \tag{19}$$

where I($\mathbf{X}; \mathbf{Y}$) denotes the mutual information shared between the two variables, and $H(\mathbf{X})$ and $H(\mathbf{Y})$ are the entropies of the two variables, respectively. I($\mathbf{X}; \mathbf{Y}$) can be further written as:

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}), \tag{20}$$

where H($\mathbf{X}$, $\mathbf{Y}$) denotes the joint entropy of two distributions.

3 Results

3.1 Point-scale evaluations

Figure 3 shows the evaluation of air temperature at both 0.25° and 0.01° against five sources of ground-based measurements. Except for the evaluation against DHM air temperature, the GMU downscaled 0.01° air temperature generally outperforms the 0.25° one. The superiority of the 0.01° air temperature is mostly demonstrated in averaged bias and averaged RMSE improvements, but less so with respect to ubRMSE and R. For example, in the comparison against CEOP air temperature, the mean bias is improved by 32% from -4.98 K (0.25°) to -3.38 K (0.01°), and the mean RMSE is improved by 23% from 5.44 K (0.25°) to 4.17 K (0.01°). However, the mean ubRMSE is degraded slightly by 0.9% from 1.91 K (0.25°) to 1.93 K (0.01°), and the mean R (= 0.96) is the same. Figure 3 also shows the evaluation of surface pressure at both 0.25° and 0.01° against ground-based CMA measurements. The downscaled 0.01° estimate yields a perfect weighted score of 4.00 (see Table 3), which means the 0.01° surface pressure is superior to the 0.25° estimate with respect to all goodness-of-fit statistics in both accuracy and precision measures. These two evaluations together signifies the benefits of detailed adjustment of the elevation difference as air temperature and pressure are very sensitive to the change of altitude especially across highly elevated regions.

Similarly, improvements are seen in the downscaled shortwave and longwave radiation estimates in the evaluation against ground-based measurements. That is, Figure 4 shows the evaluation of incident shortwave radiation, and incident longwave radiation at both 0.25° and 0.01° against CEOP measurements. In general, the 0.01° downward longwave and shortwave radiation estimates are superior to those at 0.25° especially with respect to bias and RMSE. For example, in the comparison against CEOP downward shortwave radiation, the mean bias is improved by 30% from 12.32 W/m^2 (0.25°) to 8.61 W/m^2 (0.01°), and the mean RMSE is improved by 3% from 63.02 W/m^2 (0.25°) to 61.21 W/m^2 (0.01°). In the comparison against CEOP downward longwave radiation, the mean bias is improved by 15% from -36.87 W/m^2 (0.25°) to -31.36 W/m^2 (0.01°), and the mean RMSE is improved by 6% from 43.91 W/m^2 (0.25°) to 41.23 W/m^2 (0.01°). In addition, the improvement in the downscaled 0.01° specific humidity (relative to 0.25°) is mostly demonstrated in the mean bias (see Figure 4). That is, the mean bias is improved by 74% from -0.0011 kg/kg (0.25°) to -0.0003 kg/kg (0.01°).

Figure 4 further shows the evaluation of wind speed at both 0.25° and 0.01° against three sources of ground-based measurements. On average, the range of R is generally higher (relative to other meteorological fields) possibly due to the uncertainty in wind speed measurements and estimates caused by random or turbulent disturbance, especially over the complex terrain. Generally, the 0.01° wind speed estimate slightly degrades the 0.25° result. That is, the 0.01° wind speed estimate only outperforms the 0.25° estimate in the evaluation against CMA ground-based measurements; the 0.25° wind speed estimate demonstrates better skills in the evaluation against WU or CEOP measurements. The degradations seen in the 0.01° wind speed estimates may be partly caused by the assumptions of the logarithmic wind profile used in the downscaling procedure (Rouf et al., 2019).

Table 3 summarized the weighted scores obtained from 0.01° and 0.25° near-surface atmospheric forcings estimates, respectively. It is encouraging to see that the hyper-resolution modeling improves the skill in meteorological forcing estimates (exclude precipitation) by 9% relative to coarse-resolution results. The hyper-resolution modeling outperforms the coarse-resolution meteorological forcing estimates (exclude precipitation) in nine out of 12 sets of evaluation sources in terms of estimates accuracy and precision.

Figure 5 shows the evaluation of the precipitation field used in all experiments, including HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS. It is not surprising to see that the bias-corrected CHIRPS precipitation field used in the HMA-corr-CHIRPS experiment yields a much higher positive bias compared to the rest of the precipitation estimates. This phenomenon is especially notable in the evaluation against CMA ground-based measurements in that the difference between the mean bias of precipitation estimates obtained from the HMA-corr-CHIRPS experiment at 0.01° is statistically different (at a significance level of 5%) from those obtained from all other three sets of experiments. As a result, the bias-corrected CHIRPS yields the lowest skill in precipitation estimate according to Table 4. Beck et al. (2020) argued that the disagreement between bias-corrected CHIRPS and gauge observations might be attributed to either 1) gauge under-catch issues or 2) scale mismatch between the model estimates and the gauge observations, which is reasonable. In general, the range of R is high and the mean value of R is low across all four sets of precipitation fields. The precipitation estimate skill varies more significantly over high elevated regions, whereas in flatter regions, four sets of precipitation fields demonstrate comparable skills. Comparatively, HMA-Coarse achieves the highest skills over relatively flat regions (i.e., with a mean elevation of less than 250 m). That is, the aggregated precipitation field used in the HMA-Coarse experiment at a spatial resolution of 0.25° yields a perfect score of 4.0 in the evaluation against precipitation measurements obtained from one WU station at an elevation of 250.0m. In relatively high elevations, the downscaled GMU precipitation at 0.01° yields the highest skill among all, followed by the CHIRPS precipitation at 0.01°.

Figure 6 shows the evaluation of net shortwave radiation, and net longwave radiation generated by all experiments, including HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS in the comparison against CEOP measurements. It is encouraging to see that all 0.01° net shortwave radiation estimates (obtained from HMA-GMU or HMA-CHIRPS or HMA-corr-CHIRPS) generally outperform the 0.25° estimate obtained from HMA-Coarse, especially in terms of the mean bias. For example, the mean bias is improved from 38.11 W/m^2 (HMA-Coarse) to -1.21 W/m^2 (HMA-GMU). Similarly, it is encouraging to see all 0.01° net longwave radiation estimates outperform the 0.25° estimate. The superiority of the 0.01° net longwave radiation is mostly demonstrated in averaged bias and averaged RMSE improvements, but less so with respect to ubRMSE and R. For example, the mean bias is improved by 39% from -34.80 W/m^2 (HMA-Coarse) to -21.38 W/m^2 (HMA-corr-CHIRPS), and the mean RMSE is improved by 13% from 47.33 W/m^2 (HMA-Coarse) to 41.27 W/m^2 (HMA-corr-CHIRPS). However, both of the mean R and mean ubRMSE are comparable between HMA-Coarse and HMA-corr-CHIRPS. In general, HMA-CHIRPS yields the best performance in net shortwave and net longwave radiation estimates, followed by HMA-GMU.

Figure 6 further shows the evaluation of snow depth generated by all experiments in the comparison against three sources of ground-based stations. Due to the positive bias seen within the bias-corrected CHIRPS precipitation, it is not surprising to see that HMA-corr-CHIRPS yields the worst performance due to the relatively high estimate of the snow depth relative to other experiments. For example, the mean bias is degraded from -0.05 m in HMA-GMU (or -0.06 m in HMA-CHIRPS) to 0.32 m in HMA-corr-CHIRPS. The mean RMSE is degraded from 0.33 m in HMA-GMU (or 0.29 m in HMA-CHIRPS) to 0.56 m in HMA-corr-CHIRPS. Further, the ubRMSE is degraded by 54% from 0.24 m (HMA-GMU) to 0.37 m (HMA-corr-CHIPRS). The ubRMSE is degraded by 60% from 0.23 m (HMA-CHIRPS) to 0.37 m (HMA-corr-CHIPRS). Again, it is difficult to discern whether such bad performance seen in HMA-corr-CHIRPS is due to the erroneous model estimate itself or under-representative and erroneous ground-based measurements or both. Based on the sum of the weighted scores, HMA-GMU yields the highest skill in snow depth estimates, followed by HMA-CHIRPS.

Figure 6 also shows the evaluation of skin temperature generated by all experiments in the comparison against two sources of ground-based stations. It is encouraging to see that all experiments yield relatively good agreement with the ground-based measurements in terms of R, with all Rs being greater than 0.9. All 0.01° estimates tend to correct the positive bias in the 0.25° skin temperature likely arising from the positive bias in the net shortwave radiation. That is, in the evaluation against CMA skin temperature measurements, the bias decreases from 1.16 K (HMA-Coarse) to 0.03 K (HMA-GMU), and to 0.0009 K (HMA-CHIRPS), and to -0.17 K (HMA-corr-CHIRPS). In the evaluation against CEOP skin temperature measurements, the bias drops from 1.13 K (HMA-Coarse) to -1.04 K (HMA-GMU), and to -1.06 K (HMA-CHIRPS), and to -1.47 K (HMA-corr-CHIRPS). HMA-corr-CHIRPS seems to over-correct the 0.25° skin temperature possibly due to the over-corrected precipitation, which yields the worst performance among all experiments. Although HMA-Coarse yields relatively high magnitude of the mean bias relative to both HMA-GMU and HMA-CHIRPS, HMA-Coarse yields the best performance among all experiments according to Table 4 mainly due to its superiority in the relatively low values of interquartile range (IQR; see Appendix A) achieved across all goodness-of-fit statistics.

Figure 7 shows the evaluation of soil temperature at different depths generated by all experiments in the comparison against five sets of ground-based stations. Due to the difficulty in in-situ soil temperature measurements as well as discrepancies in the measurement and model estimate depth in soil, it is not surprising to see that different experiments are superior with respect to different set of ground-based measurements. In the evaluation against CTP-SMTMN soil temperature measurements, HMA-Coarse outperforms all 0.01° estimates with respect to all goodness-of-fit statistics. Although there are 63 CTP-SMTMN stations used for evaluation, only 12 model grid cells at a spatial resolution of 0.25° are used due to the close proximity of the ground-based stations. That is, because multiple stations are colocated within one 0.25° grid cell, we evaluate the same set of 0.25° model estimates against different in-situ measurements colocated within the model grid cell. Under such circumstances, HMA-Coarse still yields the best performance partly due to relatively low spatial variability in soil temperature measurements. For example, for three 0.25° model grid cells, all with more than five colocated ground-based stations, the temporally-averaged standard deviations of the ground-based measurements are 1.28 K, 0.97 K, and 0.96 K. Further, in the evaluation against CEOP 3-cm soil temperature measurements, HMA-corr-CHIRPS yields the best skill, whereas HMA-Coarse yields the worst performance mainly due to the relatively high positive bias. That is, the bias of the 3-cm soil temperature estimates in HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS, are 3.05 K, 0.35 K, 0.36 K, and -0.24 K. In the evaluation against CEOP 4-cm soil temperature measurements, HMA-Coarse yields the best performance. HMA-Coarse is superior to all 0.01° estimates mainly in terms of significantly reduced bias and reduced RMSE. The degradation in the 0.01° estimates relative to 0.25° estimate might be caused by 1) errors in in-situ soil temperature measurements, or 2) overcorrection in the downscaled incident shortwave radiation and net shortwave radiation although the point-scale evaluation shows better performance in 0.01° estimates (see Figure 6). It is also possible that the relatively simple Inverse Distance Weighting 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. Further, in the evaluation against SETORS 4-cm soil temperature measurements, HMA-GMU yields a close-to-perfect score with improved performance seen across all goodness-of-fit statistics in terms of the accuracy measure. Compared with the estimates obtained from HMA-Coarse, HMA-GMU improves the bias by 54% from -9.21 K to -4.21 K. The RMSE is improved by 51% from 9.44 K to 4.61 K, the ubRMSE is improved by 9% from 2.07 K to 1.88 K, and the R is improved by 6% from 0.94 to 0.95. Finally, the evaluation against CEOP 5-cm soil temperature measurements shows that HMA-CHIRPS is slightly superior to other experiments. HMA-CHIRPS' better performance is largely attributed to its relatively low ranges of IQRs achieved across all goodness-of-fit statistics. To summarize, HMA-CHIRPS yields the best performance in soil temperature estimates, followed by HMA-GMU.

Table 4 summarizes the weighted score achieved by each of the experiment with respect to each set of the evaluation source. It is found that HMA-GMU yields the highest predictability skill in precipitation and model output states, followed by HMA-CHIRPS. Compared with HMA-Coarse, HMA-GMU improves the skill by 7%. However, HMA-corr-CHIRPS yields the lowest skill, which degrades HMA-Coarse predictability by 10%. These analysis, on one hand, further corroborate the importance of employing the hyper-resolution modeling versus coarse-resolution modeling strategy across the complex terrain; on the other hand, emphasize the importance of the accuracy of the hyper-resolution precipitation product used to drive model simulations.

3.2 Basin-scale evaluations

Figure 8 shows the total runoff time series obtained from all experiments for the five gauged basins in the evaluation against ground-based measurements. In general, all experiments yield relatively good agreement with the ground-based measurements in terms of both low flow and high flow seasons, except for Basin #4. In Basin #4, HMA-Coarse yields the lowest R of 0.07, and HMA-corr-CHIRPS yields the highest R of 0.66. In addition, all experiments yield positive NSEs except for Basin #3 and Basin #4. That is, HMA-corr-CHIRPS is the only experiment with a positive NSE of 0.32 for Basin #3. In Basin #4, although HMA-CHIRPS achieves the highest NSE of -0.62 among all experiments, a negative NSE is still not desirable. There can be several reasons contributing to the relative poor performance of the modeled runoff simulations in Basin #3 and Basin #4. For example, in addition to the shortcoming of neglecting water travel time (residence time) within the basin, this study does not model human-related impacts (e.g., water engineering works) and agriculture related activities (e.g., irrigation) in the total runoff simulation. Further, the discharge regime is strongly dominated by snow and glacier melt within these two basins during summer time (see Section 2.3.5), and therefore, it is possible that modeled snow melt discharge enter the stream network too soon due to too early onset of snow melt. Therefore, in Part 2 of the future study, we will determine if a simple snow cover assimilation scheme can help with modifying the snow melt timing and further improving the runoff modeling performance in snow and glacier dominated basins.

Figure 9 shows all statistics computed for evaluating the performance of HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS in comparisons against ground-based measurements. In terms of the NSE, model runs for Basin #2, Basin #3, and Basin #4 yield relatively low values (all below 0.6) as compared with Basin #1 and Basin #5. According to Table 2, Basins #2 through #4 have mean elevations of greater than 3000 m, whereas Basin #1 has a mean elevation of 1638 m and Basin #5 has a mean elevation of 681 m. Therefore, it is likely because precipitation estimates used to force models vary more significantly over high elevated regions relative to flatter regions, which is also seen in the point-scale precipitation evaluation. In addition, in flatter regions (i.e., Basin #5), all experiments yield relatively high Rs, which are greater than 0.96. Comparatively, HMA-Coarse yields the best performance across all evaluated statistics, and HMA-corr-CHIRPS yields the worst performance. In relatively high elevated regions (i.e., Basin #1 through Basin #4), 0.01° runoff estimates obtained from HMA-GMU, HMA-CHIRPS, or HMA-corr-CHIRPS are generally superior to 0.25° runoff estimates obtained from HMA-Coarse. In Basin #1, HMA-corr-CHIRPS yields the lowest bias (= 2.4 m^3/s), lowest RMSE (= 14.5 m^3/s), and highest NSE (= 0.85), whereas HMA-Coarse yields the worst performance across all statistics. In Basin #2, HMA-corr-CHIRPS seems to over-correct the total runoff especially in years 2007 through 2012. As a result, HMA-GMU yields the best performance in total runoff in terms of the lowest RMSE (= 140.2 m^3/s), lowest ubRMSE (= 121.5 m^3/s), and highest NSE (= 0.53), whereas HMA-Coarse yields the worst performance across all statistics. In Basin #3, HMA-corr-CHIRPS significantly outperforms other experiments, with a much lower bias (= -12.8 m^3/s), lower RMSE (= 352.2 m^3/s), higher R (= 0.84), and higher NSE (= 0.32). The good performance in HMA-corr-CHIPRS derived runoff might be attributed to the relatively high correction factors as applied to the region (see Figure 1b). In Basin #4, HMA-CHIRPS yields the best performance in terms of the lowest absolute value of bias (= -81.75 m^3/s), lowest RMSE (= 194.9 m^3/s), lowest ubRMSE (= 177.7 m^3/s), and less negative value of NSE (= -0.62). The over-correction issue in HMA-corr-CHIRPS runoff can also be seen from 2005 to 2012.

Since the bias-corrected CHIRPS precipitation field is obtained through calibrating against ground-based runoff measurements, it is probable that ground-based runoff measurements used in the evaluation here are also used to calibrate the bias-corrected precipitation product. This argument might be also used to explain why HMA-corr-CHIRPS can significantly outperform all other experiments in Basin #1 and Basin #3 especially in bias. However, the over-correction issue in the bias-corrected CHIRPS field should not be neglected in Basin #2 and Basin #4. Therefore, in Part 2 of the future study, we will determine if a snow cover assimilation scheme can help HMA-corr-CHIRPS to mitigate much of the positive bias possibly caused by overly-corrected precipitation.

3.3 Domain-scale evaluations

3.3.1 Evaluation of meteorological forcings

Figure 10 summarizes the multi-year averaged daily air temperature, specific humidity, surface pressure, wind speed, incident shortwave radiation, incident longwave radiation, and total precipitation before and after being downscaled from 2003 to 2016. In general, 0.01° downscaled forcings preserve the spatially and temporally averaged values obtained from original 0.25° (or 0.05°) estimates relatively well. Based on Table 6, the computed NMIs between before and after downscaled meteorological forcing field range from 0.82 to 0.96, which indicate relatively high similarities shared between the two set of forcing fields. The lowest NMI of 0.82 is obtained from the incident shortwave radiation field evaluation, which is likely due to the introduction of multiple correction factors (i.e., clearness index, local illumination, cast-shadowing, sky obstruction, and topographic configuration; Rouf et al. (2019)) in the shortwave radiation downscaling procedure.

Figure 11 shows the spatial distribution of the annual mean total precipitation obtained from HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS. The spatially-averaged annual mean precipitation difference between HMA-Coarse and HMA-CHIRPS is -5.89 mm/yr, which is largely attributed to the spatial aggregation procedure in the precipitation field used in the 0.25° estimate. Although the spatially-averaged annual mean precipitation difference between HMA-GMU and HMA-CHIRPS is negligible (= -0.74 mm/yr), precipitation magnitudes still vary grid-by-grid between these two experiments. HMA-corr-CHIRPS yields the highest precipitation magnitude in terms of the spatially-averaged mean. For example, compared with the precipitation field used in HMA-CHIRPS, the bias-corrected CHIRPS increases the spatially-averaged annual mean precipitation by 23%, with the majority of the notable increases in the mountainous regions. Despite of the discrepancies in magnitudes among all experiments, it is encouraging to see that all four total precipitation field reveal similar patterns across HMA. For example, precipitation intensity exhibits a strong north-south gradient due to orographic effects. Specifically, along the south slope of the Himalayas, annual precipitation is relatively high due to the prevalence of the Indian monsoon. While the height and extent of the Himalayas impose a significant barrier to atmospheric circulation patterns and the northward push of water vapor is greatly limited by the Himalayan mountain chain, regions north of the orographic barriers (e.g., Tibetan Plateau) receive little precipitation throughout the year (Bookhagen & Burbank, 2010). Within the Tibetan Plateau region, there exists a gradual decrease of the annual precipitation from Southeastern Tibetan Plateau to Northwestern Tibetan Plateau. The relatively dry Northwestern Tibetan Plateau is dominated by the westerlies for almost the entire year as the center of the mean moisture contribution is concentrated toward the northwest, while the Southeastern Tibetan Plateau precipitation is more influenced by the summer monsoons as the center moves more toward the southeast (You, Min, Zhang, Pepin, & Kang, 2015; Zhang et al., 2019). Overall, generally wetter regions in Bangladesh, eastern India, and the central and eastern Ganges plains are observed in all three products assessed in this study, which is consistent with the findings from Bookhagen and Burbank (2010) and Yoon et al. (2019) using other different precipitation products.

3.3.2 Evaluation of model estimates against satellite-based products

Figure 12 shows the goodness-of-fit statistics computed for HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS in the evaluation against the CGLS SWE product from 2006 to 2016 across part of HMA above latitude 35°. It is expected that the worst agreement (i.e., relatively high magnitudes of bias, RMSE, ubRMSE, and low R) of all four experiments are colocated with relatively high elevated regions inside the Tibetan Plateau, to the south of the Kunlun Mountain relative to the north of the Mountain (a.k.a., Taklamakan dessert) due to the difference in different climate regions. Although HMA-corr-CHIRPS yields the best performance in terms of the spatially-averaged bias (= -1.23 mm) compared with the rest of the experiments due to the higher total precipitation magnitude, it still yields the worst performance in terms of RMSE (= 9.87 mm) and ubRMSE (= 9.41 mm). Among HMA-Coarse, HMA-GMU, and HMA-CHIRPS, the two 0.01° SWE estimates obtained from HMA-GMU, and HMA-CHIRPS generally outperform the 0.25° SWE estimates obtained from HMA-Coarse across all goodness-of-fit statistics. In terms of the spatially-averaged bias, both HMA-GMU and HMA-CHIRPS yield slight improvements relative to HMA-Coarse. The spatially-averaged bias is improved by 13% from -2.29 mm (HMA-Coarse) to -1.99 mm (HMA-GMU), and it is improved by 12% from -2.29 mm (HMA-Coarse) to -2.02 mm (HMA-CHIRPS). Similarly, the spatially-averaged R derived by HMA-GMU and HMA-CHIRPS are improved slightly relative to HMA-Coarse. In addition, both HMA-GMU and HMA-CHIRPS yield slight improvements in RMSE and ubRMSE relative to HMA-Coarse. Overall, HMA-GMU yields the best performance in SWE estimates in the evaluation against the CGLS SWE product, followed by HMA-CHIRPS. This finding also corroborates the results in the ground-based snow depth evaluation that HMA-GMU achieves the highest score in the snow estimates.

Figure 13 shows the goodness-of-fit statistics computed for HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS in the evaluation against the MODIS skin temperature product from 2003 to 2016 across HMA. The worst agreement (i.e., relatively high magnitudes of bias, RMSE, ubRMSE, and low R) of all four experiments are along the Himalayas. The spatially-averaged bias is negative for all four experiments, however, with noticeable positive biases present in Pakistan and Northern India along Ganges and Indus rivers, covered with cropland. As discussed in Xue et al. (2019), such positive biases are possibly attributed to the lack of irrigation related activities in the Noah-MP model, and therefore yield an overestimation of the surface temperature in this region across all experiments. Comparatively, HMA-Coarse yields the most agreement (i.e., relatively low magnitudes of bias, RMSE, and ubRMSE) with the MODIS skin temperature product among all experiments, whereas HMA-corr-CHIRPS yields the worst agreement, which is consistent with the finding obtained from ground-based skin temperature evaluation. Compared with HMA-Coarse, HMA-GMU and HMA-CHIRPS decrease the spatially and temporally averaged skin temperature by 1.10 K (from 285.30 K to 284.20 K) and 1.13 K (from 285.30 K to 284.17 K), respectively (not shown). This reduction in the skin temperature magnitude is mainly caused by the reduction in the incident shortwave radiation before and after being downscaled (see Figure 10). Since HMA-Coarse already yields a negative bias in the skin temperature in the evaluation, the reduction in the HMA-GMU or HMA-CHIRPS derived skin temperature magnitude further exacerbates the negative bias, which leads to significant degradations in terms of both bias and RMSE. HMA-corr-CHIRPS skin temperature yields more negative bias than HMA-GMU and HMA-CHIRPS because more precipitation is associated with more chances of evapotranspiration, which will lead to further reduction in the skin temperature estimates. In Part 2 of the future study, we will determine if a freeze/thaw assimilation scheme can help improving the performance of the 0.01° skin temperature estimates.

4 Conclusions and discussions

This first article of a two-part series focuses on demonstrating the predictability of a hyper-resolution, offline terrestrial modeling system used for High Mountain Asia (HMA) region. To this end, this study systematically evaluates four sets of model simulations obtained from different spatial resolutions including 0.01° ($\sim$ 1-km) and 0.25° ($\sim$ 25-km) at point-scale, basin-scale, and domain-scale. The advantages of employing a hyper-resolution modeling unit (versus the coarse-resolution modeling unit) within the Noah-MP model are demonstrated in this study, especially in terms of its ability in reducing systematic errors in model estimates. That is, over relatively complex terrain, the 0.01° modeling demonstrates superiority in estimating air temperature, surface pressure, incident shortwave radiation, incident longwave radiation, specific humidity, precipitation, surface net shortwave radiation, surface net longwave radiation, snow depth, and total runoff based on point-scale and basin-scale evaluations. In terms of wind speed, skin temperature, and near-surface soil temperature, mixed performance – sometimes improvements and sometimes degradations – are seen in 0.01° estimates relative to 0.25° estimates. The exact reason of the mixed performance seen in 0.01° estimates remains unclear, but may be partly attributed to measurement errors arising from scale mismatch or measurement height discrepancies.

In the domain-scale evaluations against satellite-based products, HMA-GMU yields the largest agreement with the CGLS SWE product, and HMA-Coarse yields the largest agreement with the MODIS skin temperature product. We are aware that skill metrics computed during these comparisons are impacted by errors in the reference products. For example, the CGLS SWE product may yield higher uncertainty in estimating relatively deep snow especially over the forested regions. The accuracy of the MODIS skin temperature product is largely impacted by atmospheric attenuation effect, surface emissivity variability, as well as the procedure to derive the daily-averaged value. In this regard, systematic errors metrics such as bias and RMSE, may be secondary or tertiary as compared with the random errors measures such as ubRMSE. In Part II of this study, we will present the effects of the joint assimilation of satellite-based snow cover and freeze/thaw observations into the system. We will present to what extent the assimilation procedure will improve or degrade the performance of the 0.01° estimate without assimilation, especially for the random error measure metrics, such as ubRMSE. It is also hopeful that some of the over-correction issues seen in HMA-corr-CHIRPS can be reduced by the assimilation procedure.

Among all meteorological forcings used to drive land surface model simulations, precipitation is undoubtedly one of the most important fields. Through evaluating four sets of model simulations forced by different precipitation products, it is seen that the 0.01° estimate forced by an inaccurate precipitation representation would lead to modest degradations in model estimates relative to the 0.25° estimate. Among all 0.01° estimates, in general, HMA-GMU and HMA-CHIRPS yield relatively high skills in model estimates. Key conclusions drawn from this study are summarized below:

1) In the evaluation against ground-based measurements of air temperature, surface pressure, wind speed, incident shortwave radiation, incident longwave radiation, and specific humidity, it is found that the hyper-resolution modeling improves the skill in meteorological forcing estimates (exclude precipitation) by 9% relative to coarse-resolution estimates using the sum of the weighted scores as the criteria (see Table 3). The hyper-resolution modeling outperforms the coarse-resolution meteorological forcing estimates (exclude precipitation) in 9 out of 12 sets of evaluation sources in terms of estimates accuracy and precision. In terms of precipitation, the downscaled GMU precipitation yields the highest skill across relatively high elevated regions, which improves the predictability skill by 3% relative to the 0.25° aggregated precipitation across the complex terrain.

2) In the evaluation against ground-based net shortwave radiation measurements, all 0.01° estimates generally outperform the 0.25° estimate obtained from HMA-Coarse, especially in terms of bias and RMSE. Compared with HMA-Coarse performance in net radiation estimates, HMA-CHIRPS improves the skill by 10%.

3) In the evaluation against ground-based snow depth measurements, HMA-GMU yields the highest skill in snow depth estimates, followed by HMA-CHIRPS. Compared

with HMA-Coarse performance in snow depth estimates, HMA-GMU improves the skill significantly by 39%.

4) In the evaluation against ground-based skin temperature measurements, although HMA-Coarse yields relatively high magnitude of the mean bias relative to both HMA-GMU and HMA-CHIRPS, HMA-Coarse yields the best performance among all experiments mainly due to its superiority in the relatively low ranges of IQRs achieved across all goodness-of-fit statistics. Overall, HMA-CHIRPS degrades HMA-Coarse skill in skin temperature estimates slightly by 6%.

5) In the evaluation against ground-based near-surface soil temperature measurements, different experiments demonstrate their superiority with respect to different set of ground-based measurements. In general, compared with HMA-Coarse performance in soil temperature estimates, HMA-CHIRPS improves the skill slightly by 6%.

6) In the evaluation against ground-based total runoff measurements obtained from five gauged basins, HMA-Coarse yields the best performance across all evaluated statistics in relatively flat regions. In relatively high elevated regions, 0.01° runoff estimates obtained from HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS are generally superior to 0.25° runoff estimates obtained from HMA-Coarse.

7) 0.01° downscaled forcings preserve the spatially and temporally averaged values obtained from original 0.25° (or 0.05°) estimates relatively well with relatively high spatial similarity.

8) In the evaluation against the CGLS SWE product, HMA-GMU yields the most agreement, followed by HMA-CHIRPS.

9) In the evaluation against the MODIS skin temperature product, HMA-Coarse yields the most agreement.

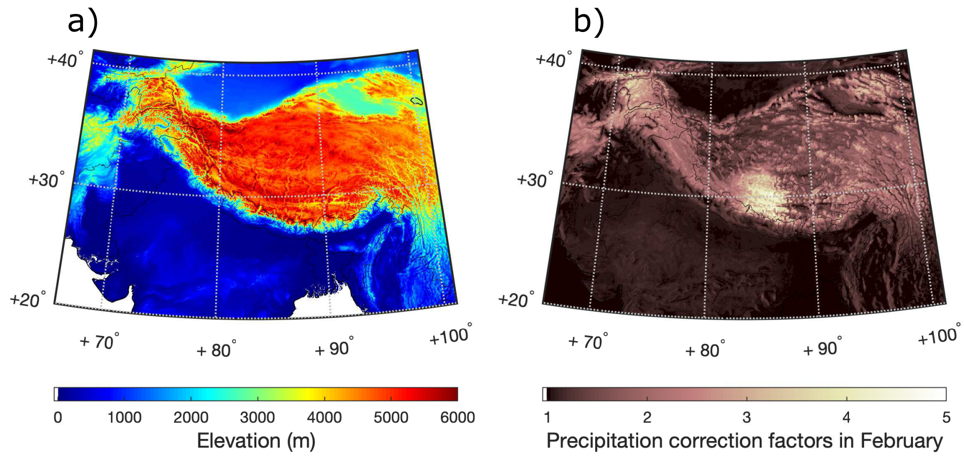

Figure 1. a) The SRTM derived HMA elevation map at a spatial resolution of 0.01°. b) An example of the spatially-distributed precipitation correction factors at a spatial resolution of 0.05° as applied in the bias-corrected CHIRPS product in February across HMA.

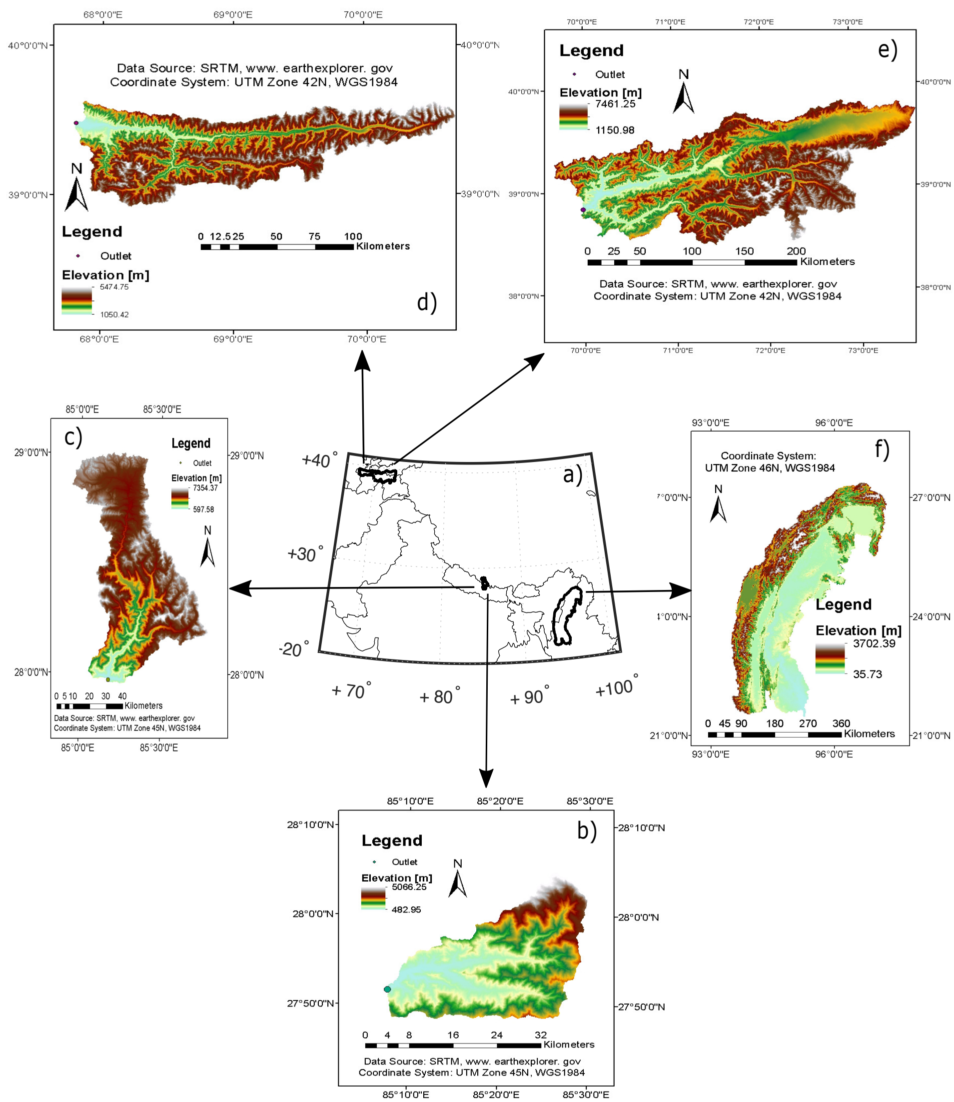

Figure 2. a) HMA study domain with gauged basin outlines in black. Gauged Basin #1 through Basin #5 are shown in b) through f) with elevation information and basin outlet locations.

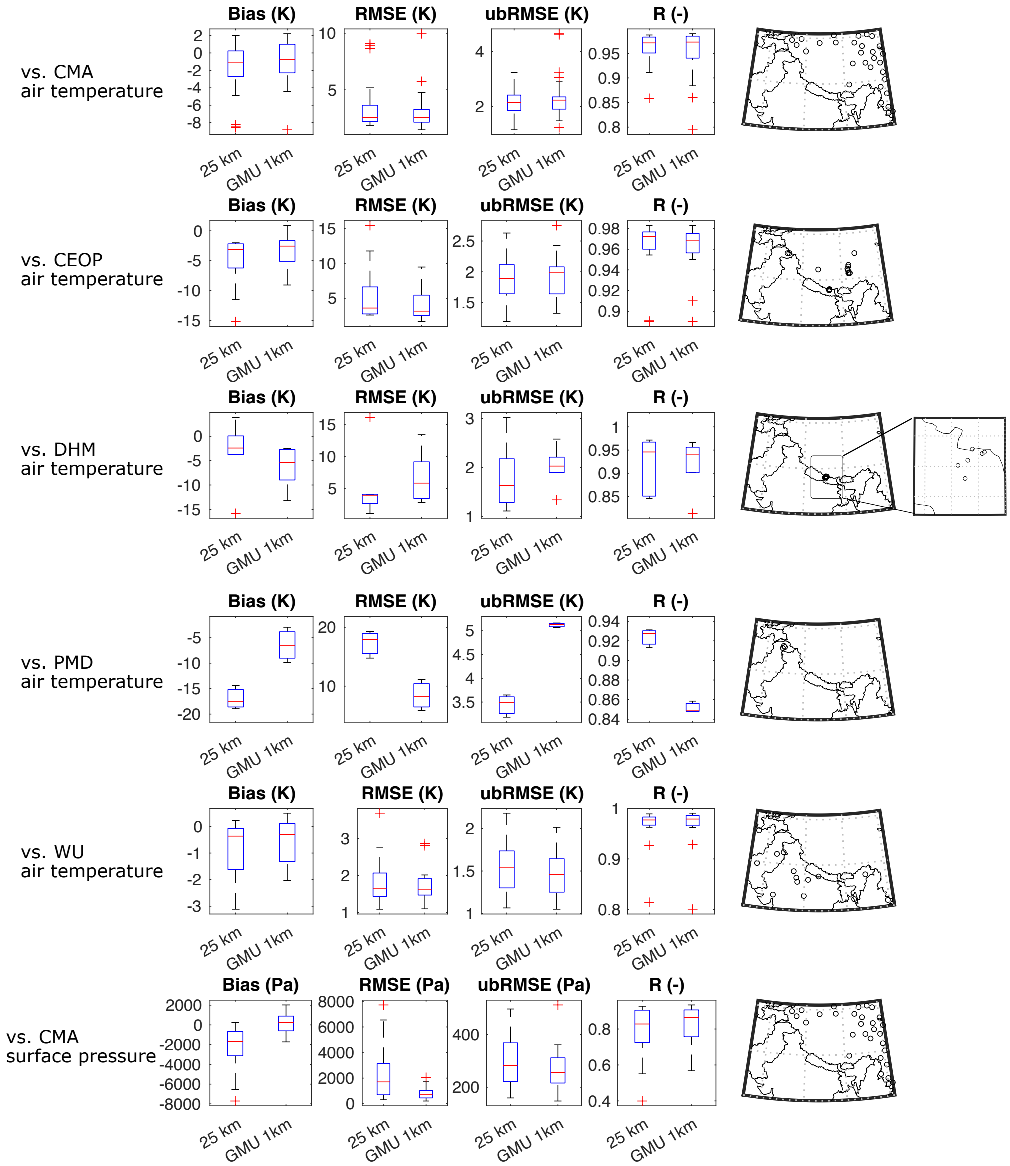

Figure 3. Box plots of bias (column 1), RMSE (column 2), ubRMSE (column 3), R (column 4) computed from 0.25° (∼25-km) and downscaled GMU 0.01° (∼1-km) meteorological forcings in the evaluation against ground-based CMA air temperature (row 1), CEOP air temperature (row 2), DHM air temperature (row 3), PMD air temperature (row 4), WU air temperature (row 5), and CMA surface pressure (row 6). The study domain with dots showing ground-based stations for each evaluation source are shown in column 5. The plus signs and red lines in the box plots are shown as outliers and medians, respectively. A close-up sub-figure of the DHM stations is shown in column 6.

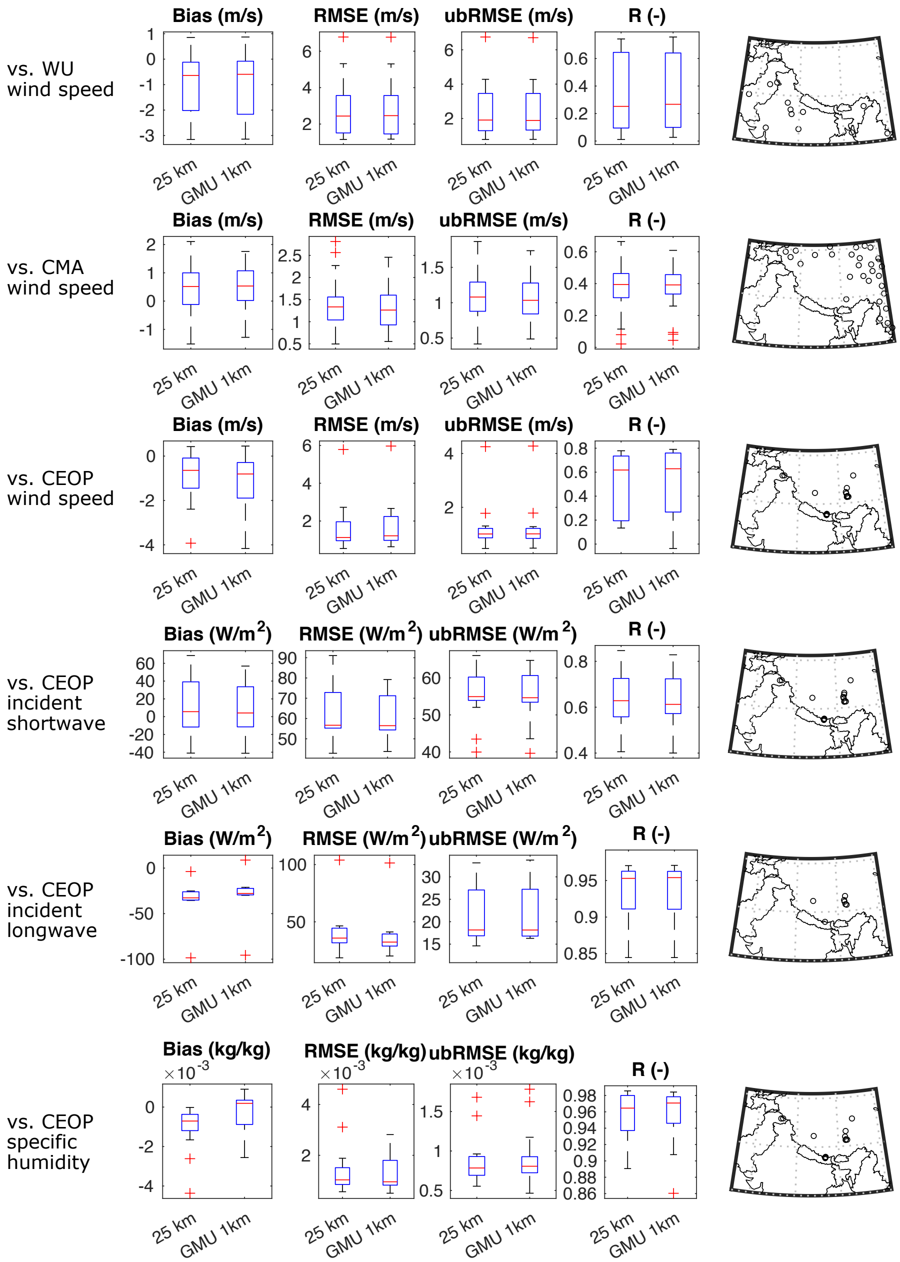

Figure 4. Same as Figure 3, but for the evaluation against ground-based WU wind speed (row 1), CMA wind speed (row 2), CEOP wind speed (row 3), CEOP incident shortwave radiation (row 4), CEOP incident longwave radiation (row 5), and CEOP specific humidity (row 6).

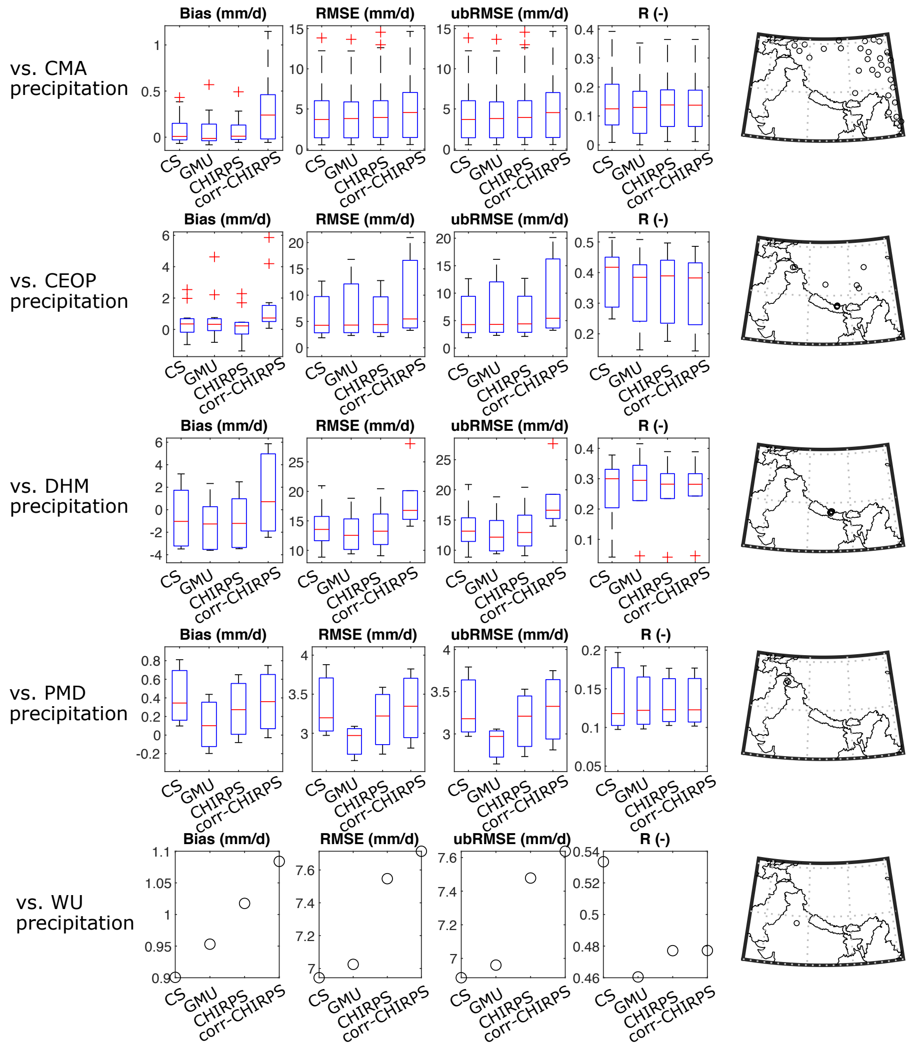

Figure 5. Box plots of bias (column 1), RMSE (column 2), ubRMSE (column 3), R (column 4) computed from HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS in the evaluation against ground-based CMA daily precipitation (row 1), CEOP daily precipitation (row 2), DHM daily precipitation (row 3), PMD daily precipitation (row 4), and WU daily precipitation (row 5). The study domain with dots showing ground-based stations for each evaluation source are shown in column 5. The plus signs and red lines in the box plots are shown as outliers and medians, respectively. The prefix of the experimental name of "HMA" is omitted for clarity.

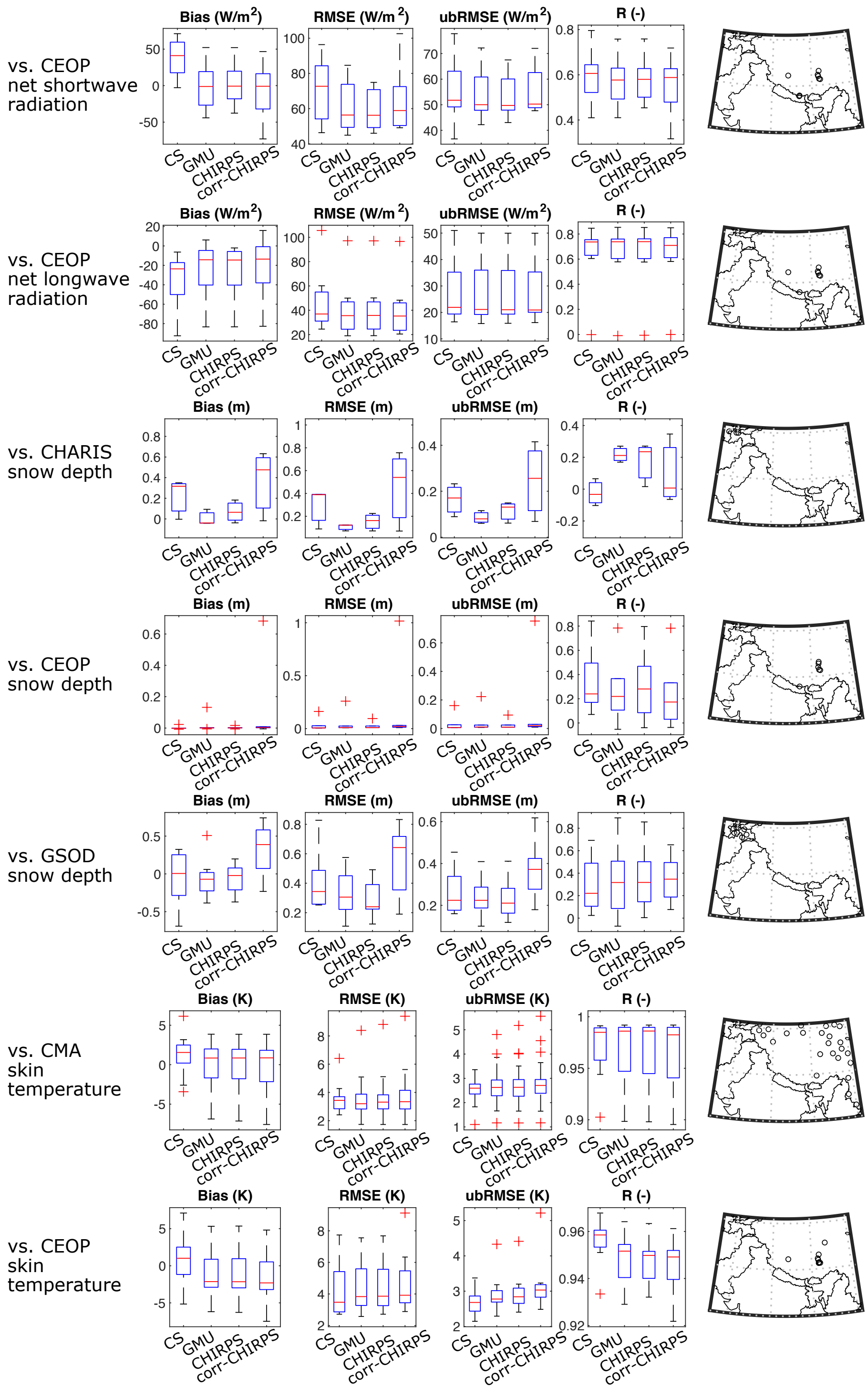

Figure 6. Same as Figure 5, but for the evaluation against ground-based CEOP net shortwave radiation (row 1), CEOP net longwave radiation (row 2), CHARIS snow depth (row 3), CEOP snow depth (row 4), GSOD snow depth (row 5), CMA skin temperature (row 6), and CEOP skin temperature (row 7).

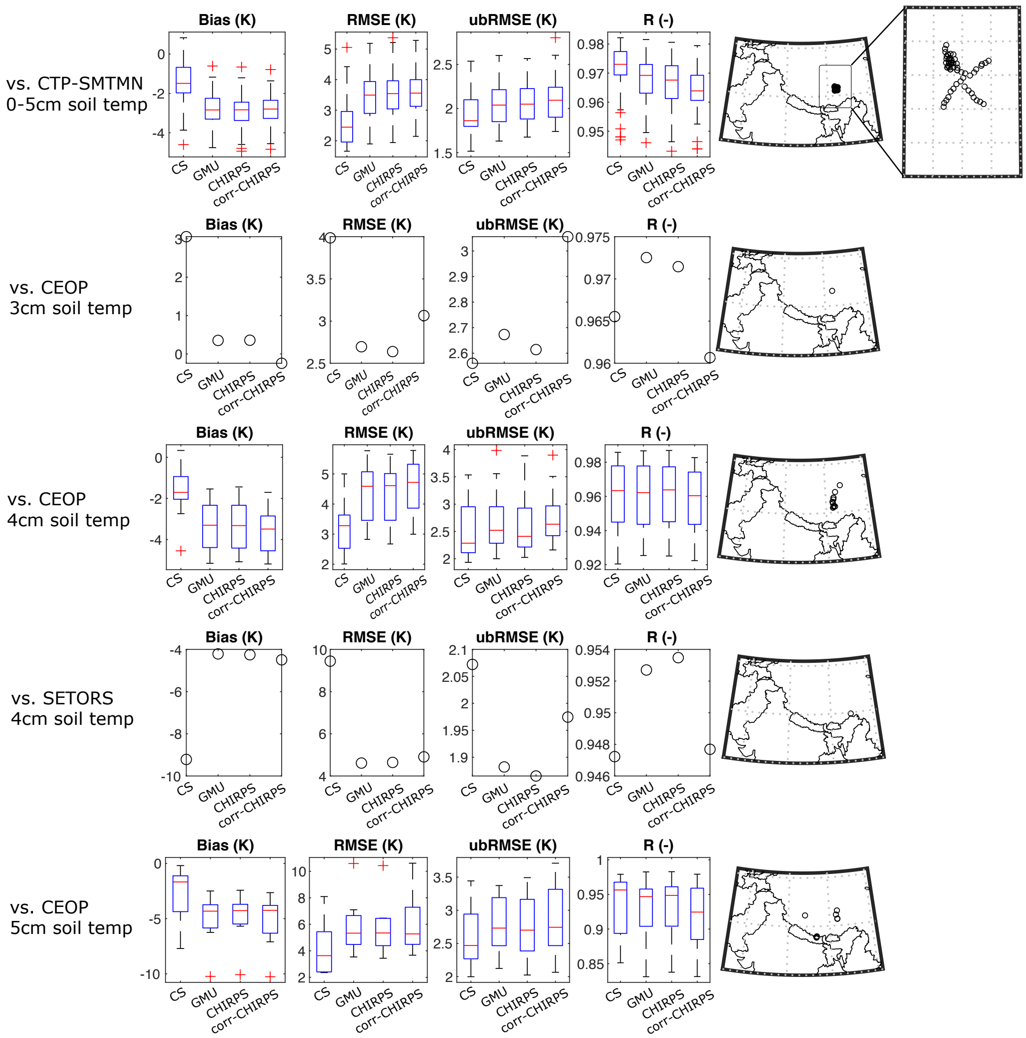

Figure 7. Same as Figure 5, but for the evaluation against ground-based CTP-SMTMN 0-5 cm soil temperature (row 1), CEOP 3 cm soil temperature (row 2), CEOP 4 cm soil temperature (row 3), SETORS 4 cm soil temperature (row 4), and CEOP 5 cm soil temperature (row 5). Note there is only one CEOP station measuring 3 cm soil temperature, and there is only one SETORS station. A close-up sub-figure of the CTP-SMTMN stations is shown in column 6.

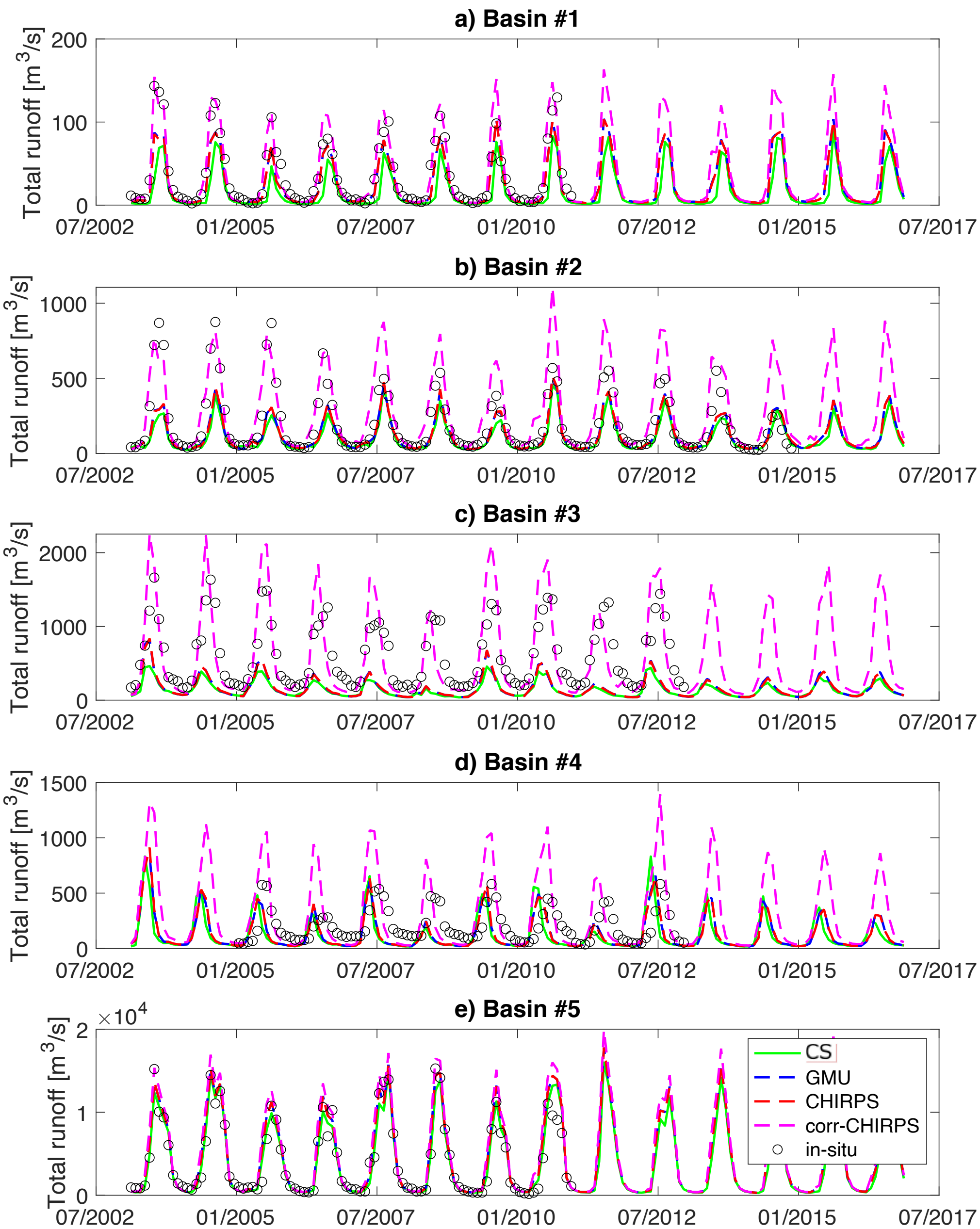

Figure 8. Monthly runoff estimates obtained from HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS for the five gauged basins in the evaluation against ground-based measurements.

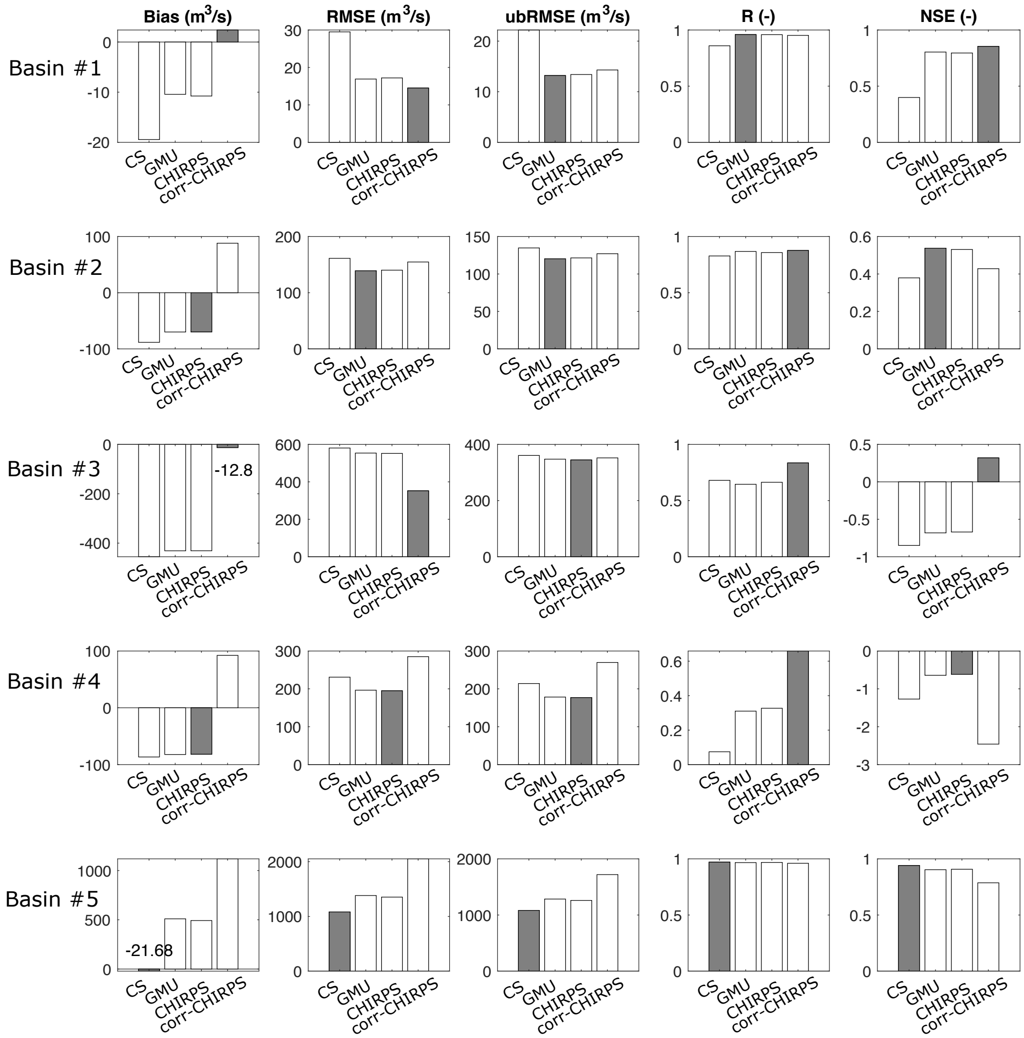

Figure 9. Statistics of bias (column 1), RMSE (column 2), ubRMSE (column 3), R (column 4), and NSE (column 5) computed from HMA-Coarse, HMA-GMU, HMA-CHIRPS, and HMA-corr-CHIRPS in the evaluation against five sets of ground-based monthly runoff measurements. Each row represents statistics for each basin. In addition, experiments with the best goodness-of-fit statistics for each basin are marked with grey bars or noted with numbers if their bars are too tiny to visualize.

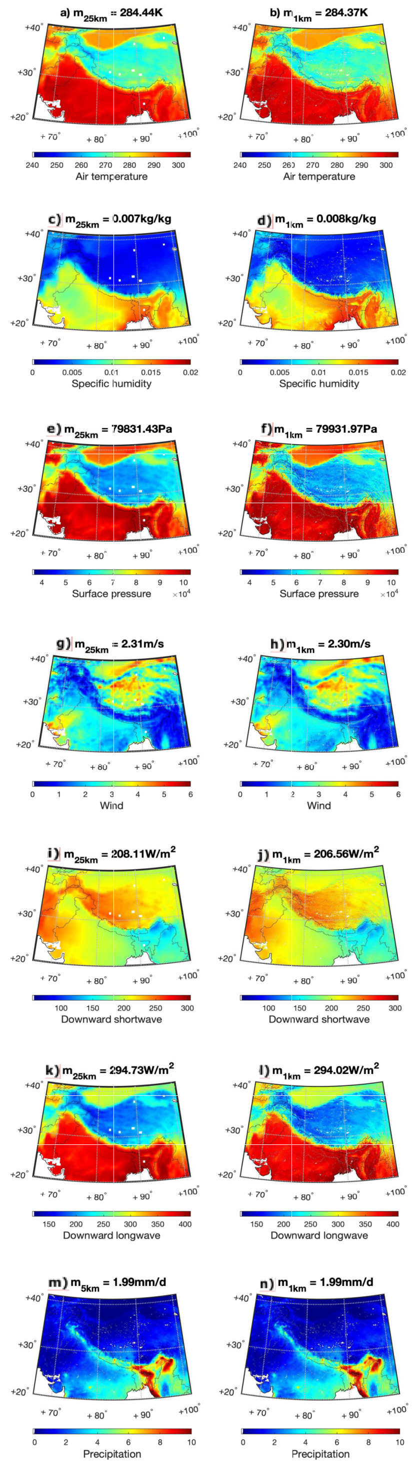

Figure 10. Multi-year (2003-2016) average of daily air temperature, specific humidity, surface pressure, wind speed, shortwave radiation, longwave radiation, and precipitation before and after being downscaled across HMA. m in the title denotes the domain-averaged value.

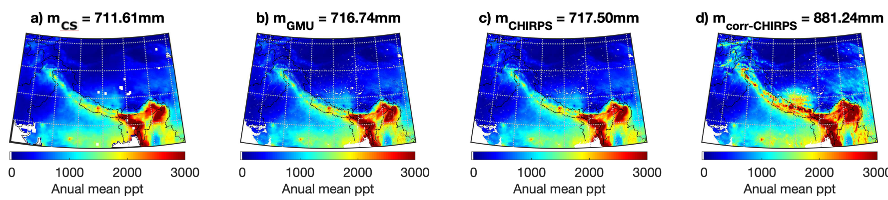

Figure 11. Annual mean total precipitation computed from a) HMA-Coarse, b) HMA-GMU, c) HMA-CHIRPS, and d) HMA-corr-CHIRPS. m in the title denotes the domain-averaged value.

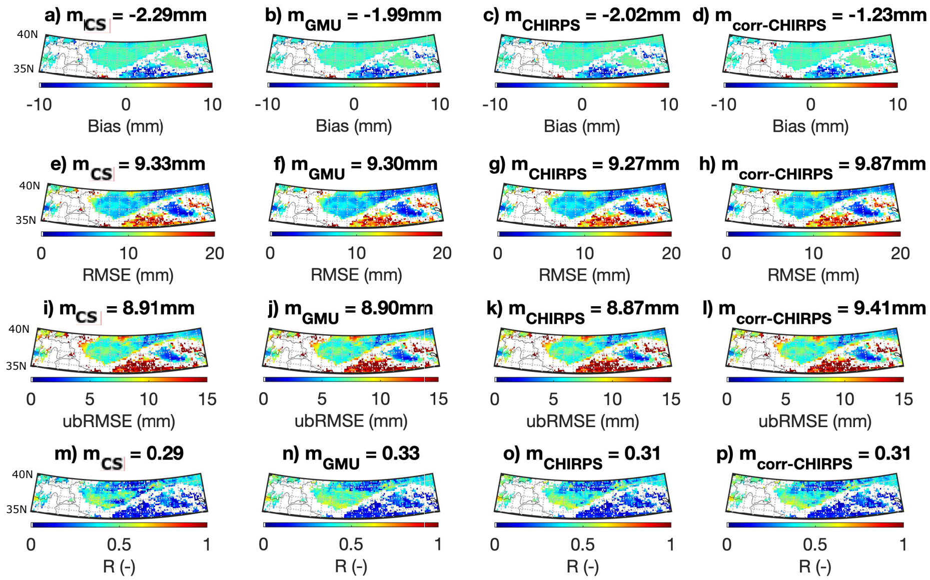

Figure 12. Goodness-of-fit statistics computed for HMA-Coarse (column 1), HMA-GMU (column 2), HMA-CHIRPS (column 3), and HMA-corr-CHIRPS (column 4) at a spatial resolution of 0.25° in the evaluation against the CGLS SWE product. Note the domain is truncated because the CGLS SWE product only covers area above latitude 35°N. Each row represents one set of goodness-of-fit statistics. m in the title denotes the domain-averaged value.

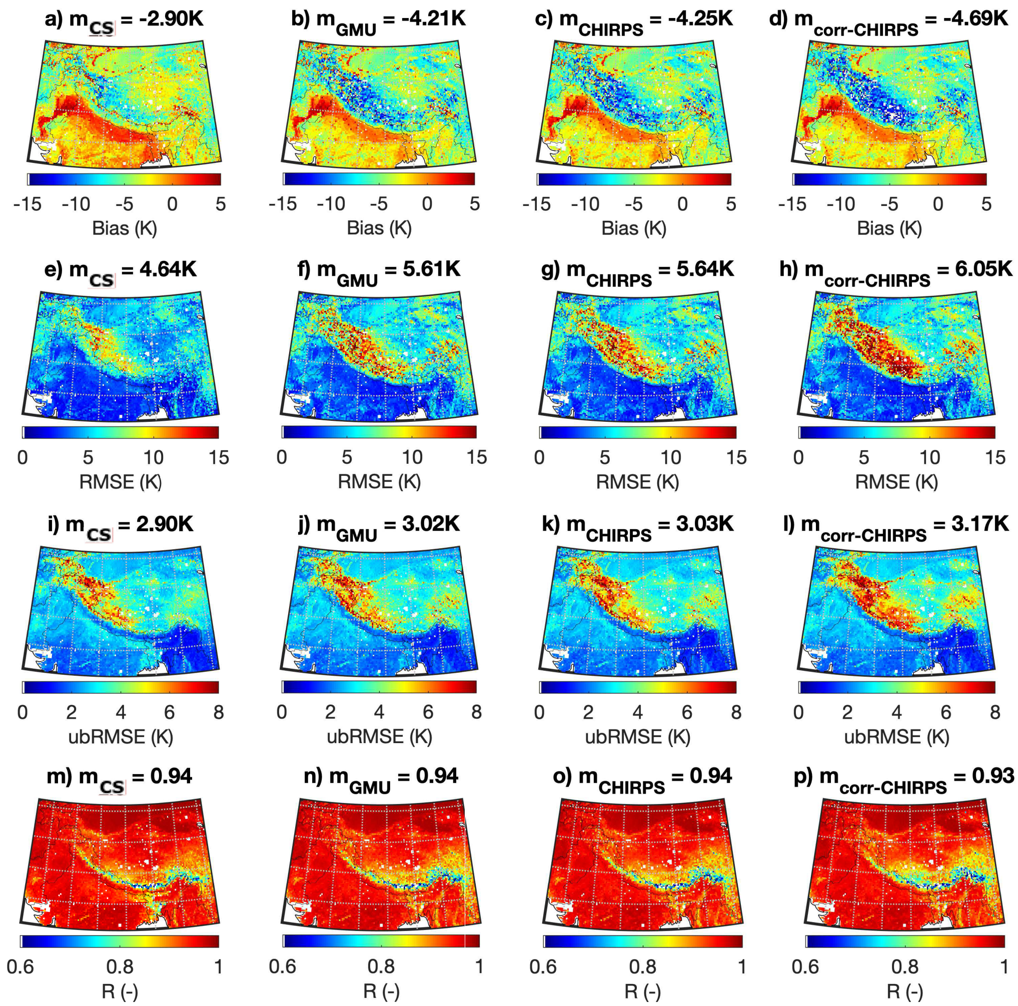

Figure 13. Goodness-of-fit statistics computed for HMA-Coarse (column 1), HMA-GMU (column 2), HMA-CHIRPS (column 3), and HMA-corr-CHIRPS (column 4) at a spatial resolution of 0.25° in the evaluation against the MODIS skin temperature product. Each row represents one set of goodness-of-fit statistics. m in the title denotes the domain-averaged value.

Table 1. Experiments used for evaluation.

Experiment name	Model output spatial resolution/temporal resolution	Precipitation input source (spatial resolution/temporal resolution)	Other meteorological forcings source (spatial resolution/temporal resolution)
HMA-Coarse (HMA-CS)	0.25°/daily	CHIRPS (0.05°/daily)	ECMWF (0.25°/6-hourly)
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)

Table 2. Summary of gauged basins shown in Figure 2. CHARIS = Contribution to High Asia Runoff from Ice and Snow project; DHM = Department of Hydrology and Meteorology in Nepal; GRDC = Global Runoff Data Centre.

Basin name (Figure number)	Drainage area (km^2)	Data Source	Mean Elevation (m)
Basin #1 (Figure 2b)	654.9	DHM	1637.9
Basin #2 (Figure 2c)	4629.1	DHM	4329.1
Basin #3 (Figure 2d)	10320.6	CHARIS	3092.8
Basin #4 (Figure 2e)	29110.9	CHARIS	3534.2
Basin #5 (Figure 2f)	110350.0	GRDC	680.7

Table 3. Summary of meteorological forcings evaluation (except for precipitation; see precipitation evaluation in Table 4) in the comparisons against ground-based stations. Forcing fields from ECMWF before downscaling at 0.25° and after downscaling at 0.01° are evaluated. The final weighted scores are calculated following the method described in Section A and higher weighted scores are bold. CMA = Chinese Meteorological Administration; CEOP = Coordinated Enhanced Observing Period project; DHM = Department of Hydrology and Meteorology in Nepal; PMD = Pakistan Meteorology Department; WU = Weather Underground.

Data Source	Number of stations (Mean elevation)	Variables (temporal scale)	Weighted score by 0.25°	Weighted score by 0.01°
CMA	30 (2442.7m)	Air temperature (daily)	3.47	**3.76**
CEOP	16 (4263.5m)	Air temperature (daily)	3.49	**3.94**
DHM	6 (2689.7m)	Air temperature (daily)	**3.41**	3.04
PMD	3 (1360.7m)	Air temperature (daily)	2.83	**3.55**
WU	15 (393.9m)	Air temperature (daily)	3.56	**3.89**
CMA	30 (2442.7m)	Surface pressure (daily)	2.29	**4.00**
WU	14 (414.1m)	Wind speed (daily)	**3.97**	3.94
CMA	30 (2442.7m)	Wind speed (daily)	3.80	**3.86**
CEOP	18 (4264.4m)	Wind speed (daily)	**3.96**	3.71
CEOP	16 (4263.5m)	Incident shortwave (daily)	3.71	**3.93**
CEOP	7 (4684.8m)	Incident longwave (daily)	3.70	**3.98**
CEOP	14 (4181.2m)	Specific humidity (daily)	3.38	**3.65**
Total scores			41.57	**45.25**

Table 4. Summary of precipitation and model states evaluation in the comparisons against ground-based stations. Experiments listed in Table 1 are evaluated. The final weighted scores are calculated following the method described in Section A and higher weighted scores are bold. CHARIS = Contribution to High Asia Runoff from Ice and Snow project; CMA = Chinese Meteorological Administration; CTP-SMTMN = Central Tibetan Plateau Soil Moisture and Temperature Monitoring Network; CEOP = Coordinated Enhanced Observing Period project; GSOD = Global Summary of the Day; SETORS = Southeastern Tibet Observation and Research Station for the Alpine Environment.

Data Source	Number of stations (Mean elevation)	Variables (temporal scale)	Weighted score by HMA-Coarse	Weighted score by HMA-GMU	Weighted score by HMA-CHIRPS	Weighted score by HMA-corr-CHIRPS
CMA	30 (2442.7m)	Precipitation (daily)	3.75	3.79	**3.83**	2.91
CEOP	11 (4036.3m)	Precipitation (daily)	3.72	3.15	**3.85**	2.49
DHM	6 (2689.7m)	Precipitation (daily)	**3.59**	3.37	3.42	3.12
PMD	3 (1360.7m)	Precipitation (daily)	2.82	**3.94**	3.05	2.86
WU	1 (250.0m)	Precipitation (daily)	**4.00**	3.79	3.62	3.53
CEOP	8 (4578.3m)	Net shortwave (daily)	3.16	3.73	**3.77**	3.22
CEOP	7 (4684.8m)	Net longwave (daily)	3.69	3.74	3.74	**3.83**
CHARIS	3 (1937.7m)	Snow depth (daily)	1.16	**4.00**	2.13	0.86
CEOP	6 (4777.9m)	Snow depth (daily)	3.38	2.87	**3.50**	1.91
GSOD	8 (2303.3m)	Snow depth (daily)	3.01	3.61	**3.71**	2.46
CMA	24 (2315.6m)	Skin temp (daily)	**3.50**	2.87	3.32	2.69
CEOP	11 (4587.3m)	Skin temp (daily)	**3.74**	3.60	3.50	3.45
CTP-SMTMN	63 (4648.3m)	0-5cm soil temp (daily)	**3.79**	3.26	3.32	3.46
CEOP	1 (5038.6m)	3cm soil temp (daily)	2.73	3.63	3.66	**3.69**
CEOP	12 (4688.5m)	4cm soil temp (daily)	**3.79**	3.05	3.07	3.24
SETORS	1 (3326.0m)	4cm soil temp (daily)	2.84	**3.99**	3.98	3.82
CEOP	9 (4356.2m)	5cm soil temp (daily)	3.48	3.46	**3.53**	3.05
Total scores			56.15	**59.85**	59.00	50.59

Table 5. Summary of reference satellite-based products used for evaluation. MODIS = Moderate Resolution Imaging Spectroradiometer; CGLS = Copernicus Global Land Service.

Data Source	Temporal coverage	Variables (temporal scale)
MODIS	01 Feb 2003 - 30 Nov 2016	Skin temperature (daily)
CGLS	01 Jan 2006 - 30 Nov 2016	SWE (daily)

Table 6. The normalized mutual information (NMI) index computed between 25-km and 1-km multi-year (2003-2016) average of daily forcing estimates (except precipitation), as well as between 5-km and 1-km multi-year average of daily precipitation estimates as shown in Figure 10.

Forcing field	NMI (-)
Air temperature	0.89
Specific humidity	0.95
Surface pressure	0.89
Wind speed	0.96
Downward surface shortwave radiation	0.82
Downward surface longwave radiation	0.93
Precipitation	0.93

A A scoring system for point-scale evaluations

Many evaluation data sources provide more than one station to compare against (see Tables 3 and 4). Therefore, the mean and the range (or spread) of the goodness-of-fit statistics (including bias, RMSE, ubRMSE, and R) are computed as measures for estimates accuracy and precision, respectively. The range of each set of goodness-of-fit statistics is calculated as the difference between the third quartile and the first quartile (a.k.a., interquartile range (IQR)). The lower the IQR is, the lower the spread is, and the higher the precision is achieved by the corresponding experiment. However, if the number of stations used for evaluation is less than three, the IQRs of goodness-of-fit statistics are not calculated, and only the means of them are calculated. As a second step, for each set of the goodness-of-fit statistics, we normalize the value (either mean or IQR of the goodness-of-fit statistics) with respect to the best statistics obtained across all experiments. Then, for each set of the model estimate, we sum up the normalized scores across all four goodness-of-fit statistics for its accuracy (mean) and precision (IQR) measures, respectively. Third, we give equal weight (50% vs. 50%) to the accuracy and the precision measures to derive the weighted score. Note that in the absence of the precision measure when the number of stations used for evaluation being less than three, we give all weight (100%) to the accuracy measure. Finally, the experiment with the highest weighted score is deemed as the best model.

Using the CEOP air temperature evaluation as an example, through averaging the bias computed via comparing against 16 ground-based stations, the mean bias of the air temperature at 0.25° (0.01°) is -4.98 K (-3.38 K). Thus, the normalized score of the 0.25° (0.01°) air temperature estimates is 0.68 (1.00) in terms of mean bias. Similarly, the IQR of bias of the air temperature at 0.25° (0.01°) is 4.04 K (3.46 K). Thus, the normalized score of 0.25° (0.01°) air temperature estimates is 0.85 (1.00) in terms of the bias IQR. Similar steps were also taken for other goodness-of-fit statistics. Then, the sum of the normalized scores in the mean of the goodness-of-fit statistics for air temperature at 0.25° (0.01°) is 3.44 (3.99). The sum of the normalized scores in the IQRs of the goodness-of-fit statistics for air temperature at 0.25° (0.01°) is 3.54 (3.89). Finally we give equal weight (50% vs. 50%) to the accuracy and the precision measures. As a result, in the evaluation against CEOP air temperature measurements, the weighted score for air temperature at 0.25° (0.01°) is 3.49 (3.94). Since the downscaled air temperature yields a higher weighted score than the original air temperature, we deem that the downscaled air temperature performs better than the air temperature at the coarse spatial resolution.

Acknowledgments

This work was funded by NASA High Mountain Asia (NNH15ZDA001N-HMA) program, grant number NNX16AQ89G. 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 WU data were obtained from `https://www.wunderground.com`. 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`. The CEOP data were obtained from `https://www.eol.ucar.edu/projects/ceop/dm/insitu/sites/ceop_ap/`. The CTP-SMTMN data were obtained from `http://dam.itpcas.ac.cn/rs/?q=data`, which was provided by Data Assimilation and Modeling Center for Tibetan Multi-spheres, Institude of Tibetan Plateau Research, Chinese Academy of Sciences. The SETORS data were obtained from `http://en.tpedatabase.cn/portal/MetaDataInfo.jsp?MetaDataId=197`. The SRTM elevation data may be downloaded from `http://srtm.csi.cgiar.org/srtmdata/`. 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. LIS models were 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 and model output generated in this study will be available at NASA Distributed Active Archive Center at National Snow and Ice Data Center.

References

Armstrong, R. L., Rittger, K., Brodzik, M. J., Racoviteanu, A., Barrett, A. P., Khalsa, S.-J. S., ... others (2019). Runoff from glacier ice and seasonal snow in high asia: separating melt water sources in river flow. *Regional Environmental Change*, *19*(5), 1249–1261.

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.

Bohn, T. J., & Vivoni, E. R. (2019). Mod-lsp, modis-based parameters for hydrologic modeling of north american land cover change. *Scientific data*, *6*(1), 1–13.

Bookhagen, B., & Burbank, D. W. (2010). Toward a complete himalayan hydrological budget: Spatiotemporal distribution of snowmelt and rainfall and their impact on river discharge. *Journal of Geophysical Research: Earth Surface*, *115*(F3).

Buck, A. L. (1981). New equations for computing vapor pressure and enhancement factor. *Journal of applied meteorology*, *20*(12), 1527–1532.

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., ... others (2003). Real-time and retrospective forcing in the north american land data assimilation system (nldas) project. *Journal of Geophysical Research: Atmospheres*, *108*(D22).

Cover, T. M., & Thomas, J. A. (1991). Entropy, relative entropy and mutual information. *Elements of information theory*, *2*, 1–55.

Dandekhya, S., England, M., Ghate, R., Goodrich, C., Nepal, S., Prakash, A., ... Udas, P. (2017). The gandaki basin: Maintaining livelihoods in the face of landslides, floods, and drought. *HI-AWARE Working Paper*, *9*.

Fiddes, J., & Gruber, S. (2014). Toposcale v. 1.0: downscaling gridded climate data in complex terrain. *Geoscientific Model Development*, *7*(1), 387–405.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., ... others (2015). The climate hazards infrared precipitation with stationsa new environmental record for monitoring extremes. *Scientific data*, *2*, 150066.

Gafurov, A., Vorogushyn, S., Farinotti, D., Duethmann, D., Merkushkin, A., &

Merz, B. (2015). Snow-cover reconstruction methodology for mountainous regions based on historic in situ observations and recent remote sensing data. *The Cryosphere*, *9*(2), 451–463.

Ghatak, D., Zaitchik, B., Kumar, S., Matin, M., Bajracharya, B., Hain, C., & Anderson, M. (2018). Influence of precipitation forcing uncertainty on hydrological simulations with the nasa south asia land data assimilation system. *Hydrology*, *5*(4), 57.

Grin, E., Schaller, M., & Ehlers, T. A. (2018). Spatial distribution of cosmogenic 10be derived denudation rates between the western tian shan and northern pamir, tajikistan. *Geomorphology*, *321*, 1–15.

Gupta, A. S., & Tarboton, D. G. (2016). A tool for downscaling weather data from large-grid reanalysis products to finer spatial scales for distributed hydrological applications. *Environmental Modelling & Software*, *84*, 50–69.

Hannah, D. M., Kansakar, S. R., Gerrard, A., & Rees, G. (2005). Flow regimes of himalayan rivers of nepal: nature and spatial patterns. *Journal of Hydrology*, *308*(1-4), 18–32.

Immerzeel, W. W., Droogers, P., De Jong, S., & Bierkens, M. (2009). Large-scale monitoring of snow cover and runoff simulation in himalayan river basins using remote sensing. *Remote sensing of Environment*, *113*(1), 40–49.

Kollet, S. J., Maxwell, R. M., Woodward, C. S., Smith, S., Vanderborght, J., Vereecken, H., & Simmer, C. (2010). Proof of concept of regional scale hydrologic simulations at hydrologic resolution utilizing massively parallel computer resources. *Water resources research*, *46*(4).

Konzelmann, T., van de Wal, R. S., Greuell, W., Bintanja, R., Henneken, E. A., & Abe-Ouchi, A. (1994). Parameterization of global and longwave incoming radiation for the greenland ice sheet. *Global and Planetary change*, *9*(1-2), 143–164.

Kulmatov, R., Opp, C., Groll, M., & Kulmatova, D. (2013). Assessment of water quality of the trans-boundary zarafshan river in the territory of uzbekistan. *Journal of Water Resource and Protection*, *5*(01), 17.

Kumar, S. V., Peters-Lidard, C. D., Mocko, D., & Tian, Y. (2013). Multiscale evaluation of the improvements in surface snow simulation through terrain adjustments to radiation. *Journal of Hydrometeorology*, *14*(1), 220–232.

Kumar, S. V., Peters-Lidard, C. D., Tian, Y., Houser, P. R., Geiger, J., Olden, S., . . . others (2006). Land information system: An interoperable framework for high resolution land surface modeling. *Environmental modelling & software*, *21*(10), 1402–1415.

Latt, Z. Z. (2015). *Flood assessment and improving flood forecasting for a monsoon dominated river basin: With emphasis on black-box models and gis* (Unpublished doctoral dissertation). Universitätsbibliothek der Leuphana Universität Lüneburg.

Lawrence, M. G. (2005). The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. *Bulletin of the American Meteorological Society*, *86*(2), 225–234.

Marshall, J., & Plumb, R. A. (1989). *Atmosphere, ocean and climate dynamics: an introductory text* (Vol. 43). Academic Press.

Mei, Y., Maggioni, V., Houser, P., Xue, Y., & Rouf, T. (2020). A nonparametric statistical technique for spatial downscaling of precipitation over high mountain asia. *Water Resources Research (Accepted)*.

Mishra, S. K., Hayse, J., Veselka, T., Yan, E., Kayastha, R. B., LaGory, K., . . . Steiner, N. (2018). An integrated assessment approach for estimating the economic impacts of climate change on river systems: An application to hydropower and fisheries in a himalayan river, trishuli. *Environmental science & policy*, *87*, 102–111.

Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T. (1996). The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, *122*(529), 73–119.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part ia discussion of principles. *Journal of hydrology*, *10*(3), 282–290.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., . . . others (2011). The community noah land surface model with multiparameterization options (noah-mp): 1. model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, *116*(D12).

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

Rouf, T., Mei, Y., Maggioni, V., Houser, P., & Noonan, M. (2019). A physically-based atmospheric variables downscaling technique. *Journal of Hydrometeorology*(2019).

Ruiz-Arias, J., Alsamamra, H., Tovar-Pescador, J., & Pozo-Vázquez, D. (2010). Proposal of a regressive model for the hourly diffuse solar radiation under all sky conditions. *Energy Conversion and Management*, *51*(5), 881–893.

Singh, R., Reager, J., Miller, N., & Famiglietti, J. (2015). Toward hyper-resolution land-surface modeling: The effects of fine-scale topography and soil texture on clm 4.0 simulations over the s outhwestern us. *Water Resources Research*, *51*(4), 2648–2667.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, *3*(Dec), 583–617.

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

Tao, J., & Barros, A. P. (2018). Multi-year atmospheric forcing datasets for hydrologic modeling in regions of complex terrain–methodology and evaluation over the integrated precipitation and hydrology experiment 2014 domain. *Journal of hydrology*, *567*, 824–842.

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

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.

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., … others (2011). The community noah land surface model with multiparameterization options (noah-mp): 2. evaluation over global river basins. *Journal of Geophysical Research: Atmospheres*, *116*(D12).

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.

You, Q., Min, J., Zhang, W., Pepin, N., & Kang, S. (2015). Comparison of multiple datasets with gridded precipitation observations over the tibetan plateau. *Climate Dynamics*, *45*(3-4), 791–806.

Yuan, F., Zhang, L., Win, K., Ren, L., Zhao, C., Zhu, Y., … Liu, Y. (2017). Assessment of gpm and trmm multi-satellite precipitation products in streamflow simulations in a data-sparse mountainous watershed in myanmar. *Remote Sensing*, *9*(3), 302.

Zhang, C., Tang, Q., Chen, D., van der Ent, R. J., Liu, X., Li, W., & Haile, G. G. (2019). Moisture source changes contributed to different precipitation changes over the northern and southern tibetan plateau. *Journal of Hydrometeorology*, *20*(2), 217–229.

Zhao, W., & Li, A. (2015). A review on land surface processes modelling over complex terrain. *Advances in Meteorology*, *2015*.