




Distributional Validation of Precipitation Data Products with Spatially Varying Mixture Models

Lynsie R. WARR , Matthew J. HEATON, William F. CHRISTENSEN, Philip A. WHITE, and Summer B. RUPPER

The high mountain regions of Asia contain more glacial ice than anywhere on the planet outside of the polar regions. Because of the large population living in the Indus watershed region who are reliant on melt from these glaciers for fresh water, understanding the factors that affect glacial melt along with the impacts of climate change on the region is important for managing these natural resources. While there are multiple climate data products (e.g., reanalysis and global climate models) available to study the impact of climate change on this region, each product will have a different amount of skill in projecting a given climate variable, such as precipitation. In this research, we develop a spatially varying mixture model to compare the distribution of precipitation in the High Mountain Asia region as produced by climate models with the corresponding distribution from in situ observations from the Asian Precipitation—Highly Resolved Observational Data Integration Towards Evaluation (APHRODITE) data product. Parameter estimation is carried out via a computationally efficient Markov chain Monte Carlo algorithm. Each of the estimated climate distributions from each climate data product is then validated against APHRODITE using a spatially varying Kullback–Leibler divergence measure.

Supplementary materials accompanying this paper appear online.

Key Words: High Mountain Asia; Latent Variables; Data Augmentation; Ordered Categorical Data.

L. R. Warr (✉)

University of California Irvine, Irvine, CA, USA.

M. J. Heaton · W. F. Christensen · P. A. White

Brigham Young University, Provo, USA (E-mail: lwarr@uci.edu).

M. J. Heaton (E-mail: mheaton@stat.byu.edu).

W. F. Christensen (E-mail: william@stat.byu.edu).

P. A. White (E-mail: pwhite@stat.byu.edu).

S. B. Rupper, The University of Utah, Salt Lake City, USA (E-mail: summer.rupper@geog.utah.edu).

© 2022 The Author(s)

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 1, Pages 99–116

<https://doi.org/10.1007/s13253-022-00515-0>

1. INTRODUCTION

1.1. PROBLEM STATEMENT AND DATA

The area known as High Mountain Asia (HMA) is comprised of several important regions, including the Indus, Ganges, and Brahmaputra watersheds. The river systems associated with each of these watersheds provide vital resources for hundreds of millions of people (Lutz et al. 2014; Zhang et al. 2019). Unfortunately, extreme events in these same watersheds also contribute to natural hazards such as flooding and landslides (Immerzeel et al. 2010; Lutz et al. 2014). Hence, added scientific understanding of these watersheds and the over 650 glaciers which feed each watershed is crucial to managing these natural resources and sustaining life in the area.

A principal driver of water availability, glacier mass balance, and glacier runoff in HMA is precipitation. Complicated by the extreme mountainous terrain, in situ observations of precipitation are sparse (Maussion et al. 2014; Palazzi et al. 2013). Hence, the primary scientific understanding of precipitation in HMA comes from digital data products such as climate models and reanalysis data—a data-assimilated combination of observations and climate modeling (Riley et al. 2018; Krishnan et al. 2019). While the value of such digital data products is immeasurable, the fact that these digital products are impacted by an incomplete understanding of the hydrological processes in HMA suggests that they are biased in their characterizations of precipitation in the region (Christensen et al. 2019; Yoon et al. 2019; Mimeau et al. 2019).

As an example, consider the following four digital data products that motivate this research. First, the Asian Precipitation—Highly Resolved Observational Data Integration Towards Evaluation (APHRODITE) data product is a continental scale data product based on statistical interpolation of rain gauge data (see <https://climatedataguide.ucar.edu/climate-data> for more information). Second, the Modern-Era Retrospective analysis for Research and Applications (MERRA-2) data product is reanalysis data based primarily on the assimilation of satellite observations with the GEOS atmospheric forecast model (see Gelaro et al. 2017, for more information). Third, the ERA5 data product is based on data assimilation of a large array of satellite, in situ and snow observations with the ECMWF weather forecast system (see <https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5> for more information). And, fourth, the Tropical Rainfall Measuring Mission (TRMM) is a purely remote sensing data product (see <https://gpm.nasa.gov/missions/trmm> for more information). Example data from each product are provided in Figure 1. Note that the products do not all have the same resolution and grid boundaries. For purposes of this research, we regridged ERA5 and MERRA-2 to be on the same grid as APHRODITE and TRMM (with $0.25^\circ \times 0.25^\circ$ squares) following the methodology of Christensen et al. (2019) in order to have matching, high-resolution grids. A future application of this research would be to explore applying this model to data products with different resolutions.

A careful inspection of Fig. 1, the mean monthly precipitation of the months April through September in the region as estimated by each data product, shows some discrepancies between the data products. For example, there is much disagreement between data products

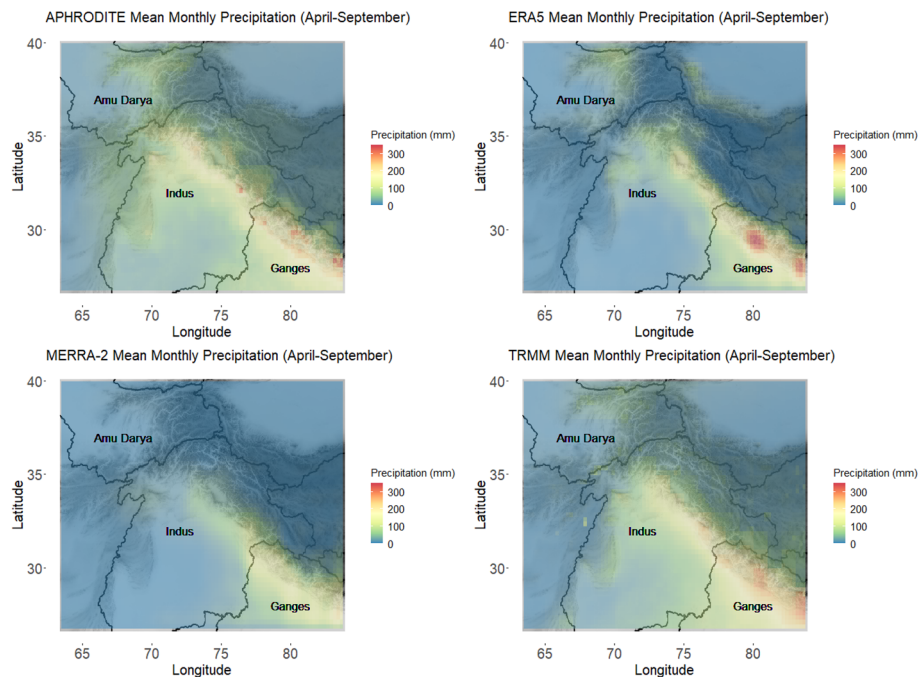


Figure 1. Mean monthly precipitation across the region of interest as estimated by each of the four data products. Watershed boundaries are outlined, and a few are labeled for reference.

about the precipitation behavior on the Himalayan crest (stretching from the center of the map to the south-east corner). ERA5 indicates some areas with much higher precipitation than the rest of the crest, but these same areas have smaller values in the APHRODITE and TRMM products and don't show up at all in the MERRA-2 product. Each product also shows varying degrees of precipitation in the areas surrounding the crest. MERRA-2 indicates that large precipitation events are essentially confined to the crest, while the other products, especially APHRODITE, show notable precipitation in various other regions. The products do agree on general trends though, and there are desert regions that all products show as having little to no precipitation.

Discrepancies such as those presented above are prevalent in digital data products. As such, data validation is required prior to using any digital data product for scientific discovery. Data validation is the process of comparing the digital data product to a "baseline" counterpart (typically observational data) to identify any potential strengths and weaknesses of the product and, potentially, correct for any systemic discrepancies. Knowing where, and how, these various data products differ allows scientists to understand where these products might be useful for scientific discovery. Validation also clearly elucidates potential biases that might enter into scientific results by using these products to, for example, inform a climate model.

While a complete review of data validation and bias correction methods for climate models is not possible here (see [Maraun 2016](#); [Chen et al. 2019](#), for holistic reviews), we briefly review the most common approaches to further motivate the contributions of this research. Data validation is most commonly done by comparing summary statistics

of the digital data product to the corresponding baseline. For example, linear scaling, or the so-called delta method, validates only discrepancies between the mean and variance of the various data products (see [Widmann et al. 2003](#); [Ratna et al. 2017](#), for examples). However, other data validation approaches include validating quantiles ([Teutschbein and Seibert 2012](#); [Jakob Themeßl et al. 2011](#)) or validating correlations amongst variables ([Vrac and Friederichs 2015](#)).

1.2. STATISTICAL CHALLENGES

While data validation is common throughout climate science, the process of data validation presents several interesting statistical challenges that are rarely addressed in the climate literature. First, the distribution of precipitation varies across space. This can easily be seen in [Fig. 2](#) which shows kernel density plots of the four data products at three different locations in the domain. These locations are shown in [Fig. 3](#). The spatial variability seen here also results in correlation between distributions at neighboring locations. The contemporary approaches to data validation mentioned above circumnavigate this problem by performing data validation one location at a time. Hence, there is a critical need for validation methods which model a smoothly changing distribution over space to account for spatial relationships.

Second, a single data product may be valid over one subregion of the spatial domain while invalid in others. Using [Fig. 1](#) as an example, the MERRA-2 data product may coincide with APHRDITE in non-mountainous areas while disagreeing with APRHODITE in mountainous regions. With spatially varying discrepancies between products, aggregated metrics (such as an overall mean) mask the true discrepancies between data products. That is, this overall metric would mask any weaknesses (and strengths) in individual data products in capturing local precipitation phenomena such as extreme events. Because of this, it is important to develop methodology that can provide an overall validation measure of the product while maintaining flexibility of validating on smaller spatial domains.

The third challenge is the matter of what to validate a data product on. For example, there are a variety of statistics that can be used for validation such as the mean, median, or quantiles that are appropriate for certain applications (see references given above). However, validating with the mean of the distribution, for example, may be problematic because means are easily influenced by tail behavior such that two similar data products could appear dissimilar due to a few very large precipitation events. While using a median would remedy this issue, such a choice would essentially ignore extremity of tail behavior, which may be scientifically important to consider.

Fourth, the data products considered here contain exact zeros (or are zero-inflated). Exact zeros coupled with positivity of precipitation do not suit the support of any standard distributions. Thus, the complexity of precipitation itself presents various modeling challenges.

1.3. RESEARCH GOALS AND CONTRIBUTIONS

In this research, we seek to implement a method for validation of these data products while accounting for the issues discussed above. Specifically, we develop a spatially varying

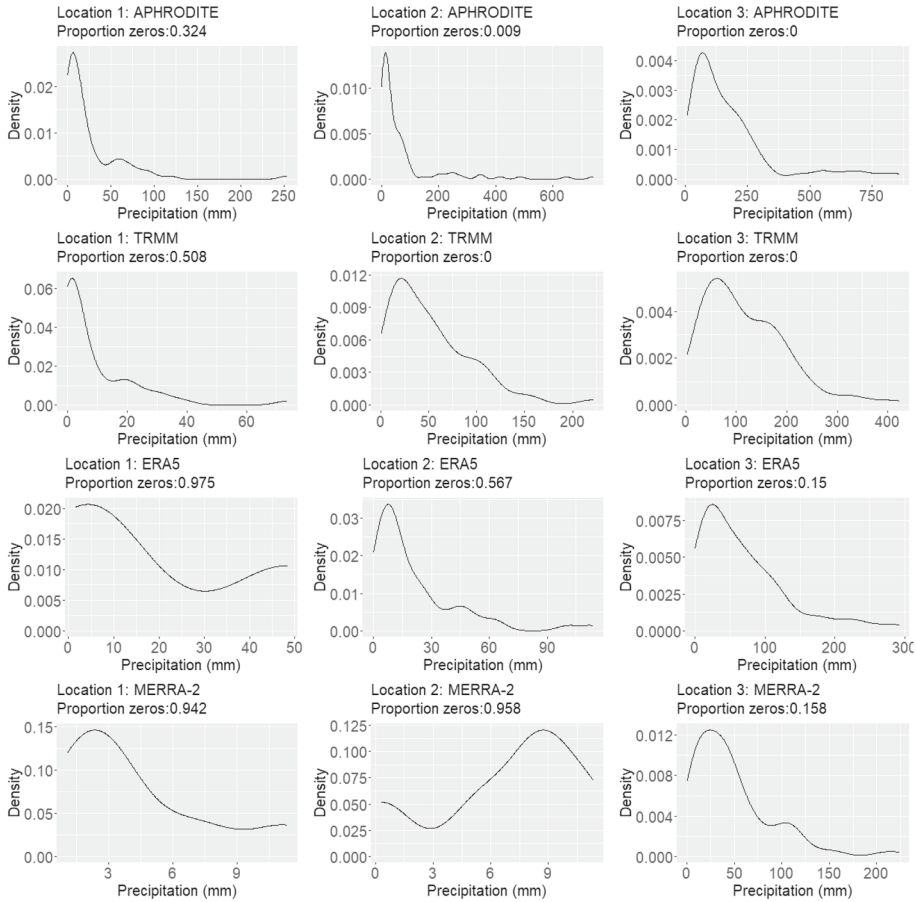


Figure 2. Density plots of mean monthly precipitation for three different locations as represented by each of the data products. (Note that zero values are excluded and summarized by a proportion.) It is clear that precipitation is represented differently by each of the data products and that these differences vary by location. Thus, it is necessary to use a spatially varying model for validation. (See Fig. 3 for a reference of these locations.).

mixture model that allows the distribution of precipitation to vary smoothly over the domain. This is accomplished by allowing the weights of the mixture model to vary smoothly over space. Importantly, by augmenting the parameter space with latent variables, we show that the majority of the parameters in our spatially varying mixture model have conjugate full conditional distributions allowing for ease of computation and good mixing of a Markov chain Monte Carlo (MCMC) sampling scheme.

Using the spatially varying mixture model, we then propose to perform validation using a pointwise Kullback–Leibler (KL) divergence measure. This pointwise KL validation metric can highlight areas where each data product is valid while also allowing aggregation across the spatial domain to produce an overall validation metric for each product. Further, because we model the entire distribution at each spatial location, our proposed methods also have the flexibility of performing validation on any summary of that distribution such as the mean, median or quantile if desired.

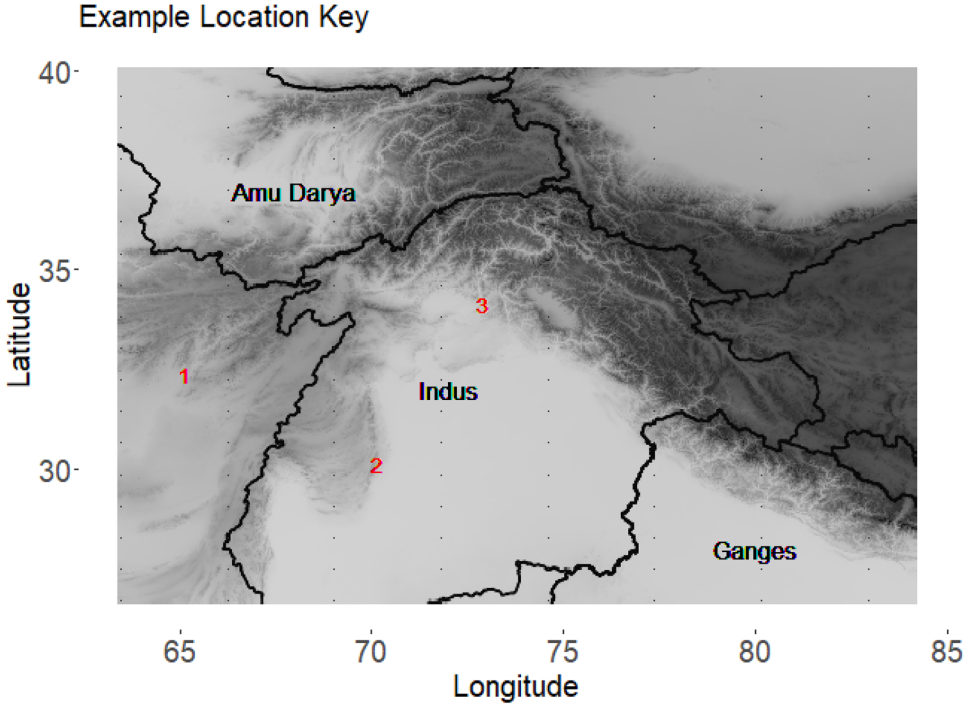


Figure 3. Locations associated with Fig. 2.

The remainder of this paper is outlined as follows: In Sect. 2, the spatially varying mixture model is presented, along with its sampling algorithm and algorithms for calculating validation metrics. In Sect. 3, the spatially varying mixture model fit results are displayed and compared using the various validation metrics. Finally, Sect. 4 contains conclusions and discusses future areas of research.

2. MODEL DESCRIPTION

In this section, we propose our spatially varying mixture model for precipitation along with our chosen metrics to perform data validation. Further, we discuss a latent variable augmentation approach that allows for more convenient posterior sampling.

2.1. SPATIALLY VARYING MIXTURE MODEL

Let $P_t(s)$ denote the precipitation value at time $t = 1 \dots T$ and location $s \in \mathcal{D}$ for a spatial domain $\mathcal{D} \subset \mathbb{R}^2$. To characterize important features of precipitation such as heavy right skewness and zero inflation, precipitation at each location was modeled as a mixture between a point mass at zero and K log-normal distributions. We assume

$$P_t(s) \stackrel{\text{ind}}{\sim} f(p|s) = \begin{cases} \omega_0(s) & \text{if } p = 0 \\ \sum_{k=1}^K \omega_k(s) \mathcal{LN}(\mu_k, \sigma_k) & \text{if } p > 0 \end{cases} \quad (1)$$

where $\{\omega_k(s)\}_{k=0}^K$ are spatially varying mixture weights and $\mathcal{LN}(\mu_k, \sigma_k)$ denotes the log-normal distribution with log-mean μ_k and log-standard deviation σ_k . Notably, while any distribution with positive support could be used, the log-normal distribution was specifically selected because of its tendency to have heavy tails, which is especially appropriate for modeling precipitation. For identifiability purposes in mixture models (Celeux 1998; Stephens 2000; Jasra et al. 2005), we order the mixture components such that $\mu_1 < \mu_2 < \dots < \mu_K$, which will also be important for modeling purposes below.

Notably, the model in Eq. (1) assumes temporal independence and no seasonal variation. Because we subset the data here to the summer season (April–September), directly accounting for seasonal variation is not necessary. Further, given that our data products are monthly precipitation, while not explicitly independent, the temporal dependence present in our data is weak and short-lived (as indicated by an exploratory data analysis). Hence, this assumption is reasonable given our considered data products for this research but may not be appropriate for all data products.

Importantly, the mixture weights in (1) are location-specific, which allows the distribution of precipitation to vary at each location. As such, some locations (e.g., the mountain ridges of HMA) may observe higher amounts of precipitation than others (e.g., the low-lying plains). In modeling the mixture weights, we desire to (i) ensure $\sum_{k=0}^K \omega_k(s) = 1$ for all s and (ii) allow $\omega_k(s)$ to vary smoothly over the spatial domain. To accomplish both of these goals, we lean on the fact that the $\{\mu_k\}$ are ordered and follow the approach of Albert and Chib (1993) by defining

$$\omega_k(s) = \int_{c_k}^{c_{k+1}} \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(\frac{-(x - \mu_u(s))^2}{2\sigma^2}\right) dx \quad (2)$$

where $c_0 = -\infty < c_1 = 0 < c_2 < \dots < c_{K+1} = \infty$ are a series of cut points and $\mu_u(s)$ is a location-specific mean. Note that, to ensure identifiability, one of the (non-infinite) cutpoints must be fixed; for our purposes, c_1 was fixed at 0. Under this parameterization, if $\mu_u(s)$ varies smoothly over space, then $\omega_k(s)$ will also vary smoothly over space. Furthermore, this parameterization allows for all $K + 1$ weights $\omega_0(s), \dots, \omega_K(s)$ to be governed by a single parameter $\mu_u(s)$, greatly reducing the parameter space.

Under (2), spatial smoothing is imposed on the $\{\omega_k(s)\}$ by imposing spatial smoothing on $\{\mu_u(s)\}$. Hence, we parameterize $\mu_u(s)$ using basis function expansions. That is, we let

$$\mu_u(s) = \mathbf{b}'(s)\boldsymbol{\theta} \quad (3)$$

where $b(s) = (1, b_1(s), \dots, b_P(s))'$ is a set of basis functions (defined below) and $\boldsymbol{\theta}$ are the associated coefficients. While any set of spatial basis functions can be used (see Cressie and Johannesson 2008; Banerjee et al. 2008; Nychka et al. 2015; Ma and Kang 2020, for examples), because this research focuses on a gridded data product, we opt to use the Moran basis functions of Hughes and Haran (2013) built from an inverse distance-weighted

neighborhood matrix. That is, for the adjacency matrix $\mathbf{A} = \{a_{ij}\}$, we set

$$a_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1/\|s_i - s_j\| & \text{if } i \neq j \end{cases} \quad (4)$$

where $\|s_i - s_j\|$ is the Euclidean distance between each pair of locations. The basis function $\mathbf{b}'(s)$ (with the exception of the intercept term) is then the s th row of the eigenvectors of the matrix $(\mathbf{I} - \mathbf{J})\mathbf{A}(\mathbf{I} - \mathbf{J})$ where \mathbf{J} is a matrix of ones. Further, the cumulative sum of the positive eigenvalues of $(\mathbf{I} - \mathbf{J})\mathbf{A}(\mathbf{I} - \mathbf{J})$ can be interpreted to represent the percentage of positive spatial variation explained by the basis functions (eigenvectors). For this research, to balance computational efficiency while still capturing spatial variability, we used $P = 94$ basis functions. These 94 basis functions accounted for, approximately, 50% of the theoretical spatial variance, but [Hughes and Haran \(2013\)](#) show that often only 10% of the spatial variation needs to be explained to adequately capture observed spatial patterns.

Under the Bayesian approach, prior assumptions were primarily selected for ease of sampling. The cutpoints $c_2 \dots c_K$ (which are all of the cutpoints that were not fixed) were transformed in order to sample more easily. These transformed cutpoints (denoted as $\delta_2 \dots \delta_K$) follow the suggestion by [Higgs and Hoeting \(2010\)](#) and are calculated as:

$$\delta_k = \log(c_k - c_{k-1}) \quad \text{where } k = 1 \dots K \quad (5)$$

The transformed cutpoints were assumed to have a uniform prior distribution. The parameter vector $\boldsymbol{\theta}$ was assumed to have a $\mathcal{N}(0, \mathbf{I})$ prior distribution. Note that this prior is somewhat informative. This is intentional: since we impose spatial smoothing on the model through $\mathbf{b}'(s)$ and $\boldsymbol{\theta}$, we enforce some level of spatial smoothing by penalizing values of $\boldsymbol{\theta}$ that are far from 0. Adjusting this prior would be one method of adjusting the strength of the spatial smoothing.

2.2. LATENT VARIABLE AUGMENTATION

The model in Sect. 2.1, while flexible, presents some computational challenges when estimating parameters. For example, the $\boldsymbol{\theta}$ parameters would require a Metropolis-type algorithm to sample from the posterior. However, in this section, we propose an equivalent model specification using latent variable augmentation that allows for more convenient posterior sampling for all parameters except for the cut points c_0, \dots, c_K .

First, let $Z_t(s) \in \{0, \dots, K\}$ represent a latent indicator for the mixture component. That is,

$$P_t(s)|(Z_t(s) = k) \stackrel{iid}{\sim} \begin{cases} \delta_0 & \text{if } k = 0 \\ \mathcal{LN}(\mu_k, \sigma_k) & \text{if } k \in \{1, \dots, K\} \end{cases} \quad (6)$$

where δ_0 is the Dirac delta function (a point mass) at 0 and $Z_t(s)$ is a discrete random variable with $\mathbb{P}\text{rob}(Z_t(s) = k) = \omega_k(s)$. Notice that marginalizing over $Z_t(s)$ yields Equation (1).

Next, because we label the mixture components based on ordering such that $\mu_1 < \mu_2 < \dots < \mu_K$, we can model $Z_t(s)$ as an *ordered* multinomial, spatial random variable. As such, we can employ the methods of [Higgs and Hoeting \(2010\)](#) and [Schliep and Hoeting \(2015\)](#) and further augment the parameter space with another latent variable $U_t(s) \sim \mathcal{N}(\mu_u(s), 1)$ such that,

$$Z_t(s) = \sum_{k=0}^K k \times \mathbb{1}\{c_k < U_t(s) < c_{k+1}\}. \tag{7}$$

Notably, integrating out the $U_t(s)$, we have $\mathbb{P}\text{rob}(Z_t(s) = k) = \omega_k(s)$ which is equivalently defined by Eq. (2).

Using the above latent variable augmentation, we can now directly sample all model parameters, with the exception of the cut points, from their conjugate complete conditional distributions. The overall Gibbs sampling algorithm is given by Algorithm 1, but here we point out a few important features of the algorithm. First, notice that there is a relationship between the $Z_t(s)$ and $U_t(s)$. That is, given $U_t(s)$, $Z_t(s)$ is known via Eq. (7). Hence, only $U_t(s)$ needs to be sampled but doing so results in a non-conjugate form of the complete conditional distribution. Therefore, Algorithm 1 samples both $Z_t(s)$ and $U_t(s)$ via composition where $U_t(s)$ is integrated out to allow for efficient sampling of $Z_t(s)$. Then, conditional on $Z_t(s)$, the complete conditional distribution of $U_t(s)$ is a truncated Gaussian distribution.

Algorithm 1: Gibbs Sampler for Spatially Varying Mixture Model

Set initial values of all parameters $\{\mu_k, \sigma_k\}$, $\{Z_t(s)\}$, $\{U_t(s)\}$, $\{c_k\}$ and θ .

for j in $1:J$ **do**

1. Sample $Z_t(s)$ and $U_t(s)$ jointly via composition by
 - a. Set $Z_t(s) = 0$ if $P_t(s) = 0$ otherwise sample $Z_t(s)$ with probabilities $\omega_k^*(s) \propto \omega_k(s) \mathcal{L}\mathcal{N}(P_t(s) \mid \mu_k, \sigma_k)$
 - b. Sample $U_t(s) \sim \mathcal{TN}(\mu_k(s), 1, c_{Z_t(s)}, c_{Z_t(s)+1})$ where $\mathcal{TN}(m, v, l, u)$ is the truncated normal distribution with mean m , variance v , lower end point l and upper end point u .
 2. Noting that $\{\log(P_t(s))\}$ where $Z_t(s) = k$ are *iid* $\mathcal{N}(\mu_k, \sigma_k)$ random variables, sample $\{\mu_k, \sigma_k\}$ via their conjugate complete conditional distribution.
 3. Sample the cut points c_2, \dots, c_K via the Metropolis accept–reject algorithm.
 4. Noting that $U_t(s)$ are *iid* $\mathcal{N}(\mathbf{b}'(s)\theta), 1)$ random variables, sample θ from its Gaussian complete conditional distribution.
-

Next, as noted in [Albert and Chib \(1993\)](#), the first cut point c_1 needs to be fixed for identifiability reasons (otherwise it is completely confounded with the mean $\mu_u(s)$). Hence, we draw c_2, \dots, c_K from their complete conditional using an adaptive Metropolis accept–reject algorithm. We again follow the convention of [Schliep and Hoeting \(2015\)](#) and integrate out the latent $U_t(s)$ to sample c_2, \dots, c_K from their posterior distribution given $\{Z_t(s)\}$ and $\mu_u(s)$.

2.3. VALIDATION METRICS

The primary validation metric used here is the Kullback–Leibler (KL) divergence measure (Kullback 1997). This metric, intuitively, measures the amount of information lost when some reference distribution $P(x)$ is approximated by another distribution $Q(x)$. Define

$$D_{KL}(P \parallel Q) = \int_{\mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) dx \quad (8)$$

to be the KL divergence where \mathcal{X} is the support of the distribution (which, in the case of our HMA application, is $[0, \infty)$). First, notice that if $P(x) = Q(x)$ for all x then $D_{KL}(P \parallel Q) = 0$ suggesting no information loss and a lower bound of zero. However, it is important to note that this metric is asymmetric, meaning that $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$, hence the necessity of choosing a reference distribution to validate against.

For the current research, we consider the KL divergence between the fitted mixture distributions in Eq. (1) for each data product and the model fit to APRHRODITE as the reference distribution. We calculate D_{KL} for each location $s \in \mathcal{D}$ using Monte Carlo integration and the fact that $D_{KL} = \mathbb{E}_P(\log(P(x)) - \log(Q(x)))$. That is, we sample precipitation values $P_t(s) \sim \hat{f}(p | s)$ where $\hat{f}(p | s)$ is the mixture distribution in Eq. (1) with parameters fixed at their respective posterior means.

We propose that digital data validation using KL divergence is most appropriate because the KL divergence captures all aspects of the distribution of the data. However, there may be specific research questions that are better addressed by comparing certain summary statistics. For example, perhaps the main quantity of interest is the extremes of the distributions in which case we may wish to validate on, say, the 0.95 quantile of the distribution. Because we focus on modeling the entire distribution of the data, we are also able to perform data validation on these other metrics. For example, to validate on a metric other than D_{KL} we can merely (i) sample many $P_t(s)$ from $\hat{f}(p | s)$ for each data product, (ii) calculate the chosen summary statistic from each of the two samples and (iii) calculate the difference of the summary statistics of each distribution. Hence, while we focus on D_{KL} , our modeling strategy is highly flexible in validating data products on various summary statistics.

3. RESULTS

The spatially varying mixture model described in Sect. 2 was fit using Algorithm 1. To assess convergence properties, three chains were run for each data product. For ERA5 and MERRA-2, 150,000 iterations were run with the first 50,000 constituting a burn-in period with the remaining being thinned by 100 to reduce autocorrelation and storage space. In our assessment of convergence, the MCMC algorithm to fit the model to APHRDITE and TRMM took longer to converge. Hence, for those data products 250,000 iterations were run, with a 50,000 burn-in period and thinning every 200th iteration. Code for implementing the sampler, as well as the data used, is available at <https://github.com/lynsiewarr/spatiallyvaryingmixture>.

In the early stages of this research, the parameters $\{\mu_k, \sigma_k\}$ were estimated as part of the MCMC algorithm outlined in Algorithm 1, but this caused the algorithm to converge

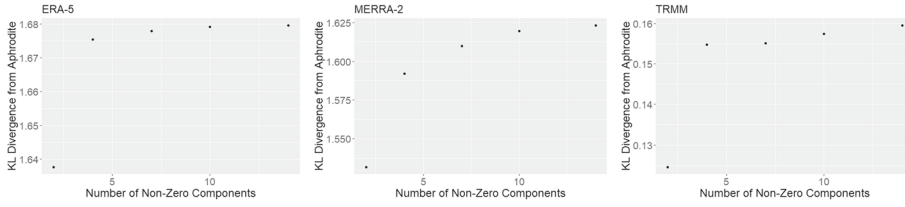


Figure 4. KL divergence from APHRODITE for the different data products for different numbers of components.

extremely slowly while creating issues with identifiability, and it necessitated handling order constraints to maintain consistent labels on the μ and σ parameters. To avoid these issues and prioritize model parsimony, the μ and σ parameters were fixed. By fixing $\{\mu_k, \sigma_k\}$ while allowing the weights $\{\omega_k(s)\}$ to be estimated, the mixture components in Eq. (1) are, effectively, a basis function expansion for the underlying distribution. This still results in a model flexible enough to capture the distribution of the data. Fixing $\{\mu_k, \sigma_k\}$ has the added benefit of improved interpretability between data products, as each part of the mixture becomes a similar “precipitation regime.” For example, it is useful to compare how much one data product utilizes the highest precipitation component to how much another utilizes it (which would not be possible if the means of the components were not consistent across data products). To select these values, the nonzero data points from all four data products were divided into K equal intervals (rather than percentiles, which would have prevented the components from accurately representing the tails) and the estimators

$$\hat{\mu} = \log \left(\frac{E[X]^2}{\sqrt{\text{Var}[X] + E[X]^2}} \right) \quad (9)$$

$$\hat{\sigma}^2 = \log \left(\frac{\text{Var}[X]}{E[X]^2} + 1 \right) \quad (10)$$

were used where $E[X]$ was calculated as the middle of the interval, and $\text{Var}[X]$ was calculated as the square of half the width of the interval.

The total number of nonzero components K was selected by examining both DIC and the KL divergence between APHRODITE and each other product, for several different numbers of components. To enable good mixing and avoid identifiability problems, it would be optimal to use as few mixture components as possible while still capturing the differences between products. The DIC values were 1.55e7, 1.39e7, 1.35e7, 1.34e7 and 1.33e7 for $K = 2, 4, 7, 10, 14$, respectively. Further, Fig. 4 shows the changing KL divergence for each product across the numbers of components. For ERA5 and TRMM, it appears that most of the differences between the product in question and APHRODITE are captured at $K = 4$, as the increase in KL divergence and decrease in DIC slow there. However, we selected $K = 7$ (for all data products) to better capture the differences between MERRA-2 and APHRODITE, since the KL divergence for MERRA-2 did not plateau as quickly and the DIC value was lower.

Convergence was assessed by examining trace plots (from one of the chains) and calculating Monte Carlo standard error (Flegal et al. 2008) and the Gelman–Rubin diagnostic

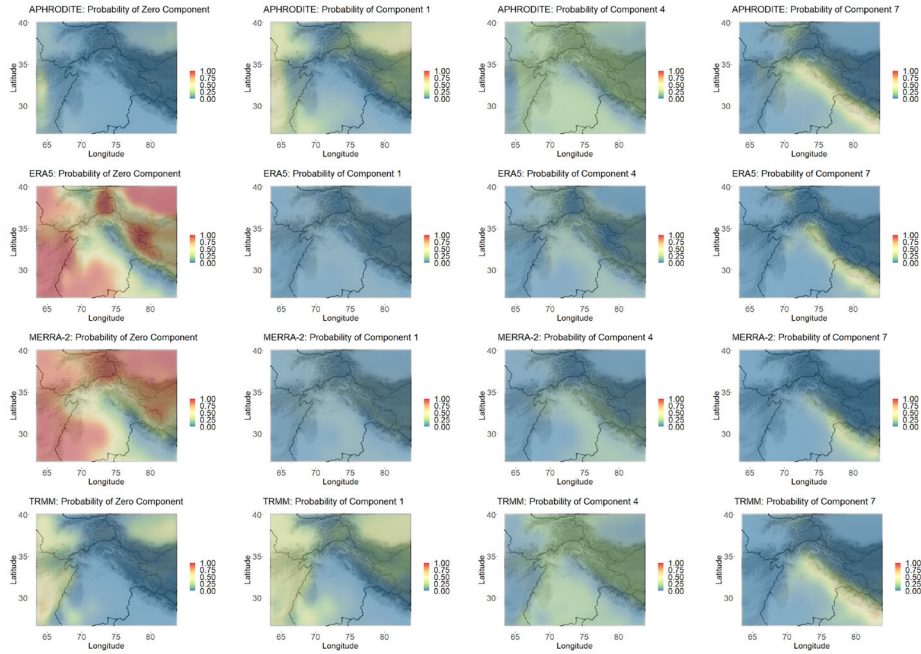


Figure 5. Mixture weights of the zero component and the first, fourth, and last lognormal components of the mixture distribution across space.

(Cowles and Carlin 1996). While some parameters mixed and converged better than others, overall the convergence diagnostics indicated that we achieved sufficient mixing.

3.1. MODEL FIT AND QUALITATIVE PRODUCT COMPARISON

The first step in performing product validation according to the above methods is ensuring that the fitted mixture distributions match the each data product individually. However, examining the model fit for the spatially varying mixture model is a daunting task since there are distributional fits for each data product across over 4000 locations in our example. Further, in many of the locations, there were very few nonzero precipitation values to assess the model fit on. Thus, to assess model fit, we first examined the difference in the estimated probability of the precipitation being zero (ω_0) compared to the actual proportion of zero values in the data across space (see Figure 1 in the supplementary material). We also examined the estimated distribution compared with a kernel density estimate of the data at a few example locations. Both of these comparisons indicated a model fit that is satisfactory for all the data products. Admittedly, however, both of these comparisons do not necessarily constitute a full model fit evaluation since our model includes spatial smoothing constraints. However, the results from these comparisons give confidence in our data validation below.

As a first qualitative validation of the data products, because we fixed the mixture components, we can compare the data products based on the probability of precipitation belonging to any given mixture component. It is clear from Fig. 5 that the model fits for the ERA5 and MERRA-2 data products heavily utilize the zero component in much of the desert regions,

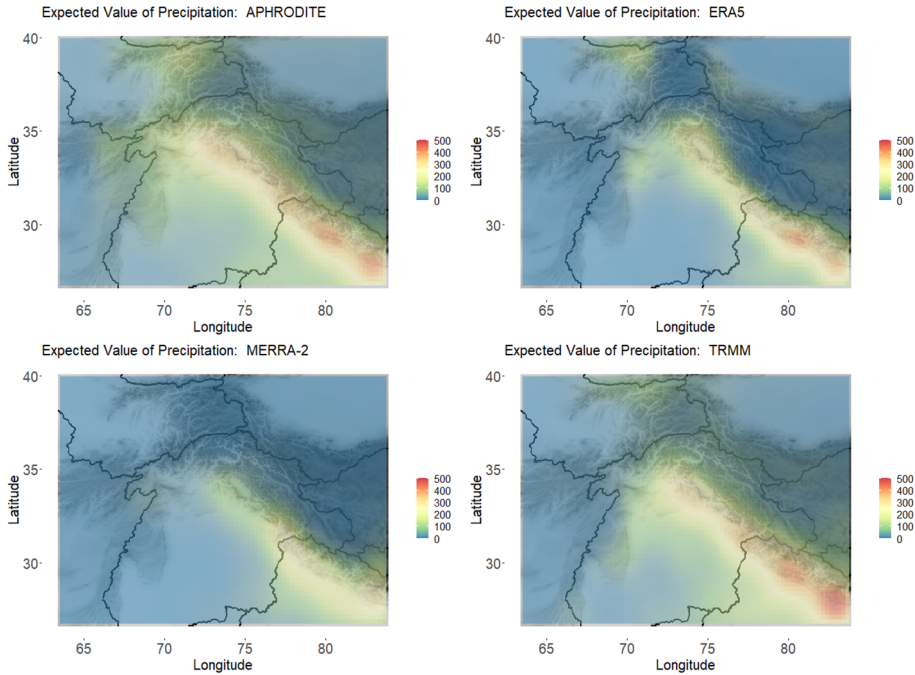


Figure 6. Expected value (mm/month) of the distributions at each location across the region.

while APHRODITE and TRMM use it to a lesser degree. APHRODITE and TRMM appear to detect (or predict) more small, nonzero precipitation events, which are captured in the small and moderate components (components 1 and 4 are shown as examples). The different data product fits all appear to use the most extreme component (component 7) to similar degrees.

To further qualitatively compare the different data products, we examined the expected value of the distributions at each location as shown in Fig. 6. These figures also seem to indicate a closer similarity between APHRODITE and TRMM than between APHRODITE and ERA5 or MERRA-2. These differences can be seen in the amount of area that have expected values close to zero, and in the higher extremes along the ridge.

3.2. QUANTITATIVE PRODUCT COMPARISON

In order to quantitatively compare ERA5, MERRA-2, and TRMM relative to APHRODITE, the Kullback–Leibler divergence was calculated between the model fit to each of the data products and the model fit to APHRODITE (the reference distribution). The KL divergence was calculated between the distributions at each individual location. The KL divergence across the region for each comparison can be seen in Fig. 7.

Consistent with the qualitative analysis in the previous section, it seems that TRMM is most similar to APHRODITE in nearly every region according to this metric (though parts of the high KL region in TRMM may be outperformed by ERA5). For ERA5 and MERRA-2, the similarities are much closer along the Himalayan crest, while the desert regions are more

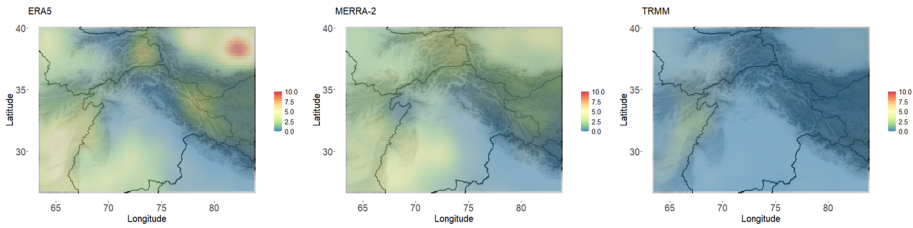


Figure 7. Kullback–Leibler divergence from APHRODITE for each of the three other data products across the region of interest. According to this metric, TRMM has the precipitation distributions closest to APHRODITE’s over most of the region .

Table 1. Average KL divergence for the overall region, region 1, and region 2 (as highlighted in Fig. 8) for each of the three data product comparisons to APHRODITE

	Overall	Region 1	Region 2
ERA5	1.678	0.194	2.240
MERRA-2	1.610	0.297	1.979
TRMM	0.155	0.018	0.385

dissimilar to APHRODITE (likely because of their inability to detect smaller precipitation events as discussed above). As an overall measure of data product validation, the average KL divergence values across the region can be seen in Table 1.

One particular advantage of our method is that it allows for comparison in certain regions of interest. For example, we can calculate the KL divergence between the model fit for APHRODITE and the model fits for the other data products specifically for the two regions highlighted in Fig. 8. The average KL divergence for each of those regions, as well as the overall KL divergence, is shown in Table 1 for each data product. Notably, each data product seems to perform the best (in terms of comparison to APHRODITE) along the Himalayan crest (Region 1). However, there is more discrepancy between the data products on the western edge of HMA (Region 2).

Our validation approach here focuses on estimating the entire distribution of precipitation across the spatial domain. However, because we estimate the distribution, we can easily perform validation of the different data products for various statistics of interest. For example, we can compare the different data products based on the mean, median, or 95th quantile of the fitted distribution. An example of such statistical validation (as opposed to distributional validation) is given in Fig. 9, which displays a spatial map of the difference in the 95th quantile (similar maps showing the difference in mean and median are included in supplementary materials).

Figure 9 shows that each data product differs quite substantially from APHRODITE in terms of extreme precipitation. That is, along the Himalayan crest, ERA5, MERRA-2 and TRMM all seem to understate extreme precipitation (compared with APHRODITE), while overstating extreme precipitation in the high plains and valleys. It is important to note that the greatest discrepancies in the 95th quantiles occur along the Himalayan crest, while the Himalayan crest has relatively small discrepancies according to the KL divergence metric.

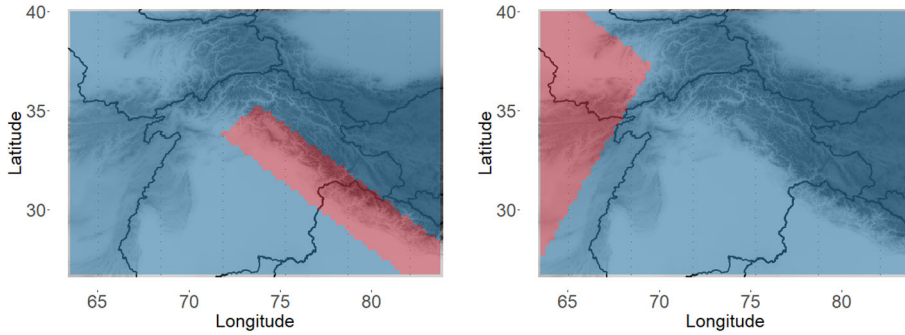


Figure 8. Highlighted regions (red) indicate regions that KL divergence is calculated for in Table 1. Region 1 is on the left, and region 2 is on the right.

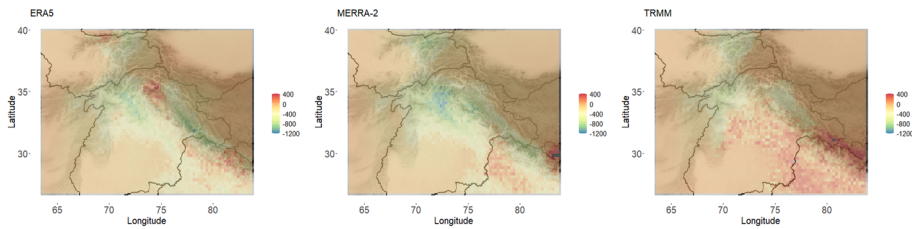


Figure 9. Difference in 95th quantile (mm/month) of precipitation distributions between APHRODITE and the other data products across the region. The difference is calculated as the 95th quantile of the fitted distribution to APHRODITE subtracted from the 95th quantile of the fitted distribution to the other data product.

This demonstrates the use of KL divergence as a general distribution summary that is not oversensitive to outliers. We would argue this is a better representation of the distribution as a whole.

4. CONCLUSION

In this research, we developed a spatially varying mixture model to estimate the density of precipitation across a heterogeneous region in High Mountain Asia. Through the use of latent variable techniques, we also developed a computationally feasible way of fitting the associated mixture to big data. Having a fitted distribution for precipitation, we then validated various precipitation data products for the region using KL divergence and other distribution summary statistics. Importantly, this validation enables either point-by-point or global comparisons of the products so as to inform scientists on the strengths and weaknesses of each product.

While this work is an interesting first validation of data products, this work does not answer *why* observed differences occur. For example, does elevation explain the difference in the distributions? An interesting follow-up analysis would be to develop some sort of regression model that explains the differences between the distributions. This, in its own right, has various statistical challenges including defining a regression model with a dif-

ference between distributions as a response. These are important questions that we hope to address in future research.

Though the various metrics for comparison reveal different information about the distribution similarities between data products, all the validation metrics used in this research indicate that TRMM precipitation is the most similar to APHRODITE precipitation in model fit at most locations. However, the approach used here allows for the identification of specific locations where another data product may be more comparable according to the preferred metric. In other words, using our modeling methods we are able to validate each data product in any given user-defined region.

We specifically recommend using the KL divergence metric for comparison as it evaluates the entire distribution and is not overly sensitive to outliers. Because of these features, we believe it provides a better picture of the water resources available (which is especially important around the Himalayan crest since that is where most of the precipitation occurs).

While the proposed mixture model generally fit the precipitation data well, the μ and σ parameters could also be estimated to potentially improve model fit. In early phases of this research, estimating μ and σ was attempted for this problem but there were issues with convergence and identifiability that made the implementation difficult. Furthermore, in terms of model fit, the locations with small amounts of precipitation were most prevalent because they far outnumber the high precipitation locations. This resulted in poor fits for the high precipitation locations, which may be unacceptable for applications where the right tail is of scientific interest.

For this research, we considered validation of data products on the same resolution. However, different data products are often available on different resolutions and grids. A potential future research avenue is to develop similar methodology that can be applied to different data products at different resolutions and grids.

ACKNOWLEDGEMENTS

This research was funded by NASA (80NSSC20K1594).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

[Received August 2021. Revised May 2022. Accepted August 2022. Published Online September 2022.]

REFERENCES

- Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. *J Am Statist Assoc* 88(422):669–679

- Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. *J R Statist Soc Ser B (Statist Methodol)* 70(4):825–848
- Celeux G (1998) Bayesian inference for mixture: the label switching problem. *Compstat. Physica, Heidelberg*, pp 227–232
- Chen J, Brissette FP, Zhang XJ, Chen H, Guo S, Zhao Y (2019) Bias correcting climate model multi-member ensembles to assess climate change impacts on hydrology. *Climat Change* 153(3):361–377
- Christensen MF, Heaton MJ, Summer Rupper C, Reese S, Christensen WF (2019) Bayesian multi-scale spatio-temporal modeling of precipitation in the indus watershed. *Front Earth Sci* 7:210
- Cowles MK, Carlin BP (1996) Markov chain monte carlo convergence diagnostics: a comparative review. *J Am Statist Assoc* 91(434):883–904
- Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. *J R Statist Soc Ser B (Statist Methodol)* 70(1):209–226
- Flegal JM, Haran M, and Jones GL (2008) Markov chain monte carlo: Can we trust the third significant figure? *Statist Sci* 23(2):250–AC260
- Gelaro R, McCarty W, Suárez MJ, Todling R, Molod A, Takacs L, Randles CA, Darmenov A, Bosilovich MG, Reichle R et al (2017) The modern-era retrospective analysis for research and applications, version 2 (merra-2). *J Climate* 30(14):5419–5454
- Higgs MD, Hoeting JA (2010) A clipped latent variable model for spatially correlated ordered categorical data. *Comput Statist Data Anal* 54(8):1999–2011
- Hughes J, Haran M (2013) Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J R Statist Soc Ser B (Statist Methodol)* 75(1):139–159
- Immerzeel WW, Van Beek LPH, Bierkens MFP (2010) Climate change will affect the Asian water towers. *Science* 328(5984):1382–1385
- Jasra A, Holmes CC, Stephens DA (2005) Markov chain monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist Sci* 20(1):50–67
- Krishnan R, Shrestha AB, Ren G, Rajbhandari R, Saeed S, Sanjay J, Syed MA, Vellore R, Xu Y, You Q et al. (2019) Unravelling climate change in the Hindu Kush Himalaya: rapid warming in the mountains and increasing extremes. In: *Hindu Kush Himalaya Assessment*, Springer, Cham, pp 57–97
- Kullback S (1997) *Information theory and statistics*. Courier Corporation, Chelmsford
- Lutz AF, Immerzeel WW, Shrestha AB, Bierkens MFP et al (2014) Consistent increase in high Asia's runoff due to increasing glacier melt and precipitation. *Nat Climate Change* 4(7):587–592
- Ma P, Kang EL (2020) A fused gaussian process model for very large spatial data. *J Comput Gr Statist* 29(3):479–489
- Maraun D (2016) Bias correcting climate change simulations—a critical review. *Curr Climate Change Reports* 2(4):211–220
- Maussion F, Scherer D, Mölg T, Collier E, Curio J, Finkelnburg R (2014) Precipitation seasonality and variability over the Tibetan plateau as resolved by the high Asia reanalysis. *J Climate* 27(5):1910–1927
- Mimeau L, Esteves M, Jacobi H-W, Zin I (2019) Evaluation of gridded and in situ precipitation datasets on modeled glacio-hydrologic response of a small glacierized himalayan catchment. *J Hydrometeorol* 20(6):1103–1121
- Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution gaussian process model for the analysis of large spatial datasets. *J Comput Graph Stat* 24(2):579–599
- Palazzi E, Von Hardenberg J, Provenzale A (2013) Precipitation in the Hindu-Kush Karakoram Himalaya: observations and future scenarios. *J Geophys Res Atmosp* 118(1):85–100
- Ratna SB, Ratnam JV, Behera SK, Tangang Fredolin T, Yamagata T (2017) Validation of the WRF regional climate model over the subregions of southeast Asia: climatology and interannual variability. *Climate Res* 71(3):263–280
- Riley C, Rupper S, Steenburgh WJ, Strong C, Kochanski A (2018) Characteristics of extreme precipitation events in high mountain Asia as inferred from high resolution regional climate modeling. *AGUFM* vol 2018, pp C21E–1386

- Schliep EM, Hoeting JA (2015) Data augmentation and parameter expansion for independent or spatially correlated ordinal data. *Comput Statist Data Anal* 90:1–14
- Stephens M (2000) Dealing with label switching in mixture models. *J R Statist Soc Ser B (Statist Methodol)* 62(4):795–809
- Teutschbein C, Seibert J (2012) Bias correction of regional climate model simulations for hydrological climate-change impact studies: review and evaluation of different methods. *J Hydrol* 456:12–29
- Themeßl MJ, Gobiet A, Leuprecht A (2011) Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int J Climatol* 31(10):1530–1544
- Vrac M, Friederichs P (2015) Multivariate-intervariable, spatial, and temporal-bias correction. *J Climate* 28(1):218–237
- Widmann M, Bretherton CS, Salathé EP (2003) Statistical precipitation downscaling over the northwestern united states using numerically simulated precipitation as a predictor. *J Climate* 16(5):799–816
- Yoon Y, Kumar SV, Forman BA, Zaitchik BF, Kwon Y, Qian Y, Rupper S, Maggioni V, Houser P, Kirschbaum D et al (2019) Evaluating the uncertainty of terrestrial water budget components over high mountain Asia. *Front Earth Sci* 7:120
- Zhang F, Thapa S, Immerzeel W, Zhang H, Lutz A (2019) Water availability on the third pole: a review. *Water Secur* 7:100033

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.