AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2017WR020635

**Key Points:**
- Assessment of similarity is important in determining a multimodel ensemble
- Similarity can provide an assessment of the model utility to an ensemble, separate from accuracy measures
- Similarity is dependent on model variable, climate regime, and time scale of interest

**Supporting Information:**
- Supporting Information S1

**Correspondence to:**
S. V. Kumar,
Sujay.V.Kumar@nasa.gov

**Citation:**
Kumar, S. V., Wang, S., Mocko, D. M., Peters-Lidard, C. D., & Xia, Y. (2017). Similarity assessment of land surface model outputs in the North American Land Data Assimilation System. *Water Resources Research*, *53*, 8941–8965. https://doi.org/10.1002/2017WR020635

Received 24 FEB 2017
Accepted 5 OCT 2017
Accepted article online 10 OCT 2017
Published online 12 NOV 2017



# Similarity Assessment of Land Surface Model Outputs in the North American Land Data Assimilation System

**Sujay V. Kumar[1], Shugong Wang[1,2], David M. Mocko[1,2], Christa D. Peters-Lidard[3], and Youlong Xia[4]**

[1]Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA, [2]Science Applications International Corporation, McLean, VA, USA, [3]Earth Science Division, NASA Goddard Space Flight Center, Greenbelt, MD, USA, [4]I.M. Systems Group at Environmental Modeling Center, NCEP, College Park, MD, USA

**Abstract** Multimodel ensembles are often used to produce ensemble mean estimates that tend to have increased simulation skill over any individual model output. If multimodel outputs are too similar, an individual LSM would add little additional information to the multimodel ensemble, whereas if the models are too dissimilar, it may be indicative of systematic errors in their formulations or configurations. The article presents a formal similarity assessment of the North American Land Data Assimilation System (NLDAS) multimodel ensemble outputs to assess their utility to the ensemble, using a confirmatory factor analysis. Outputs from four NLDAS Phase 2 models currently running in operations at NOAA/NCEP and four new/upgraded models that are under consideration for the next phase of NLDAS are employed in this study. The results show that the runoff estimates from the LSMs were most dissimilar whereas the models showed greater similarity for root zone soil moisture, snow water equivalent, and terrestrial water storage. Generally, the NLDAS operational models showed weaker association with the common factor of the ensemble and the newer versions of the LSMs showed stronger association with the common factor, with the model similarity increasing at longer time scales. Trade-offs between the similarity metrics and accuracy measures indicated that the NLDAS operational models demonstrate a larger span in the similarity-accuracy space compared to the new LSMs. The results of the article indicate that simultaneous consideration of model similarity and accuracy at the relevant time scales is necessary in the development of multimodel ensemble.

## 1. Introduction

Multimodel ensembles are often used in weather, climate, and hydrologic modeling efforts in order to quantify and understand model uncertainty (Dirmeyer et al., 2006; Mitchell et al., 2004; Murphy et al., 2004; Palmer et al., 2005; Pitman & Henderson-Sellers, 1998; Rodell et al., 2004; Xia et al., 2012b). These efforts demonstrate that combining individual model estimates generally leads to increased skill, as the individual model errors tend to cancel each other out (Xia et al., 2012a, 2012c). In addition to simple averaging, different methods of combining individual model estimates have been proposed (Giorgi & Mearns, 2002; Tebaldi et al., 2005), based on how well the predictions from each model agree with observations. An evaluation of the multimodel ensembles with and without observational constraints is often useful for evaluating the predictability limits and uncertainty sources of models (Schwalm et al., 2015). As the availability of reliable, spatially distributed, and temporally consistent observations is not always guaranteed, the use of observational constraints in ensemble modeling is not always possible. In addition, a key assumption behind the assumed improved skill of the ensemble mean estimates is that the constituent models and their model parameterizations are independent of each other. Therefore, other ways of examining the relative contributions of constituent models are necessary while working with an ensemble of models.

Similarity is an intuitive criterion for assessing things of the same kind, such as comparing outputs of an ensemble of models for a given variable. As a fundamental and widely used concept, different metrics can be used to quantify similarity. Most similarity quantifications belong to either distance-based or information-based measurements (Li et al., 2004). Distance-based similarity measures (e.g., mean squared error) are often used in applications such as cluster analysis (Tryon, 1939), whereas information-based measures (e.g., correlation) are typically used for information retrieval applications (Singhal, 2001). In order to define similarity between two quantities *A* and *B*, both commonality and differences have to be considered.

The more common two quantities are, the more similar they are. Conversely, if more differences can be found in a comparison, the less similar the compared quantities are. The maximum similarity between *A* and *B* is reached when *A* and *B* are identical (Lin, 1998). In this article, we present a similarity based assessment of a suite of land surface models (LSMs) on their contribution toward a multimodel ensemble.

The similarity evaluations are conducted using the outputs from a suite of LSMs in the North American Land Data Assimilation System (NLDAS; Mitchell et al., 2004; Xia et al., 2012b) configuration. NLDAS is a multiinstitution, real-time, and retrospective land data assimilation system that runs four LSMs using observation-based meteorological data. NLDAS Phase 2 (NLDAS-2) was implemented into the NCEP central operations (NCO) in August 2014. The operational NLDAS-2 uses Noah (version 2.8; Noah28; Chen et al., 1996a; Ek et al., 2003; Xia et al., 2012c), Mosaic (Koster & Suarez, 1992, 1996), Variable Infiltration Capacity (VIC version 4.0.3; VIC403; Liang et al., 1994), and Sacramento Soil Moisture Accounting (SAC; Burnash et al., 1973) LSMs. As part of ongoing efforts, new LSMs or new versions of the LSMs are being considered for the next phase of the NLDAS project to replace or augment the existing suite of LSMs. In this article, in addition to the four operational NLDAS-2 model outputs, we consider the model estimates from four LSMs in the NLDAS-2 configuration: Noah version 3.6 (Noah36; Wang et al., 2010; Wei et al., 2013), VIC version 4.1.2.l (VIC412L; Gao et al., 2010), Catchment version 2.5 (CLSM25; Ducharne et al., 2000; Koster et al., 2000; Reichle et al., 2011), and a configuration of NoahMP (as implemented in the Weather Research and Forecasting model V3.6) LSM with dynamic vegetation (NoahMP; Niu et al., 2011; Yang et al., 2011).

There have been several prior studies that have quantitatively compared and evaluated the outputs from the NLDAS suite of models. Model evaluation studies (Pan et al., 2003; Robock et al., 2003; Sheffield et al., 2003) during the first phase of the NLDAS project (NLDAS-1) were primarily focused on evaluating the model outputs against available reference measurements. The deficiencies in individual model formulations identified in these studies led to Phase 2 of the NLDAS project, where model parameterizations and boundary condition inputs were improved (Xia et al., 2013). Though the model evaluations conducted in NLDAS-2 (Xia et al., 2012b) indicate greater level of agreement between the constituent models relative to those in NLDAS-1, significant intermodel differences were also observed, particularly for cold season and subsurface hydrologic processes. Comparison of the land surface energy and water budgets (at monthly and annual scales) from the NLDAS-2 models was examined in more recent studies (Xia et al., 2016a, 2016b), which found similarities among the models in the simulation of seasonal cycles and significant biases relative to reference data products. An evaluation of the components of the terrestrial water storage, including snow and soil moisture, from the NoahMP and CLSM25 LSMs in the NLDAS configuration was presented in Xia et al. (2017). These studies are largely focused on evaluating the skills of the individual models and less focused on quantifying when and how the constituent model estimates are similar and what factors best explain convergence in areas with high level of model agreement.

As the primary reason for using an ensemble of models is to reduce the overall prediction uncertainty, it is important to assess the similarity across the models. If the intermodel spread is small (the models are similar to each other), then it can be argued that they are not truly independent and their utility to a model ensemble is low, irrespective of their simulation skill. Conversely, a multimodel ensemble is meaningful only if there is sufficient dissimilarity among the constituent models. For example, Mo et al. (2012) examined the uncertainties from two different LDASs and concluded that the ensemble skill and uncertainty was primarily driven by the input meteorological forcing more than the intermodel differences within a given system. Though having models that produce vastly different estimates is not helpful for reducing the overall uncertainty, outlier models may incorporate important processes or different formulations of model physics. Understanding factors of dissimilarity among the individual models is useful for choosing the models that constitute an ensemble (Rastetter, 1996). In this article, we provide quantitative assessments of similarity of a suite of LSMs within NLDAS-2 through a confirmatory factor analysis. The article also presents assessments of model similarity relative to measures of agreement between model estimates compared to observational references.

Specifically, the article addresses the following research questions and objectives:

1. How similar/dissimilar are the constituent models within the operational NLDAS-2 and the new LSMs that are considered for inclusion within NLDAS-2? When and where are the similarities between the models more/less prominent?

2. How does similarity/dissimilarity of models change at different time scales? As NLDAS-2 is used for a variety of applications (e.g., drought assessment, flood risk estimation, and weather/climate model initialization), it is important to assess if the ensemble is providing meaningful estimates of modeling uncertainty at a given time scale.
3. Can acceptable levels of model performance for constituent models be established without them becoming too similar to each other? This can be quantified through assessing trade-offs in model performance (level of agreement between model estimates and observations) and model similarity for constituent models.

The article is organized as follows: section 2 introduces the settings of this study, including the domain, models, variables of interest, and methods used for similarity assessment. The results of intermodel similarity assessment and factor analyses are described in section 3. Similarity assessments at different time scales and evaluation of model similarity in relation to accuracy are given in sections 4 and 5, respectively. In the end, the main findings of this study are summarized in section 6.

## 2. Study Settings

### 2.1. Models

As noted above, model outputs from eight LSMs are used in this article for similarity assessment. They include the four NLDAS-2 operational models (Noah28, Mosaic, VIC403, and SAC) and four additional LSMs (Noah36, CLSM25, VIC412L, and NoahMP). SAC represents the combination of Sacramento Soil Moisture Accounting (SAC) rainfall runoff model and the SNOW-17 empirical snow pack model (Anderson, 1973). Version 3.6 of the Noah land surface model (Chen et al., 1996b) represents several physics improvements and model fixes, including to snow physics (Wang et al., 2010) and to warm season processes (Wei et al., 2013) over the current operational NLDAS-2 version of the Noah LSM. The VIC model (Liang et al., 1994) version 4.1.2.l has evolved from 4.0.3, including upgrades such as new parameterizations of the soil temperature profile, snow cover, and frozen ground physics. Developed from the Noah LSM, the Noah multiphysics (NoahMP) LSM has integrated several physics modules from other land surface models. CLSM land surface model is partially developed based on the Mosaic model (Koster & Suarez, 1992) and represents the land component of the NASA Goddard Earth Observing System model version 5 (GEOS-5) system.

All LSMs except SAC employ similar physics components with different parameterizations for soil hydrology, canopy interception, soil thermodynamics, and snowpack physics, which are summarized in Table 1. Across the LSMs, the heat flow calculations are generally defined by the usual diffusion equation:

$$C(\Theta)\frac{\partial T}{\partial t}=\frac{\partial}{\partial z}\left[K_t(\Theta)\frac{\partial T}{\partial z}\right], \tag{1}$$

where $T$ is the soil temperature and the volumetric heat capacity $C$ ($\mathrm{J\,m^{-3}\,K^{-1}}$) and the thermal conductivity, $K_t$ ($\mathrm{W\,m^{-1}\,K^{-1}}$) are formulated as functions of soil moisture $\Theta$ in LSMs. For example, in Noah28, they are defined as (Chen & Dudhia, 2001)

$$C=\Theta_{water}C_{water}+(1-\Theta_s)C_{soil}+(\Theta_s-\Theta)C_{air}+(\Theta-\Theta_{water})C_{ice}, \tag{2}$$

$$K_t=\begin{cases}420\exp[-(2.7+P_f)], & P_f\le 5.1\\ 1.1744, & P_f>5.1\end{cases}, \tag{3}$$

and $P_f=\log[\Psi_s(\Theta_s/\Theta)]$, where $\Psi_s$ is the saturated water potential (suction), $\Theta_s$ is the maximum soil moisture (porosity), $\Theta_{water}$ is the unfrozen portion of soil moisture, and $C_{water}$, $C_{soil}$, $C_{air}$, and $C_{ice}$ are the volumetric heat capacities of water, soil, air, and ice, respectively. In VIC412L, the heat capacity $C$ is calculated similarly as the weighted average of these capacities, while in VIC403, frozen ground physics were not used, so the heat capacity of ice is not considered. However, the soil heat conductivity is calculated differently as

$$K_t=(K_{t,sat}-K_{t,dry})K_e+K_{t,dry} \tag{4}$$

in which, $K_e$ is the Kersten number describing the degree of saturation (Gao et al., 2010). This formulation is also used in NoahMP to calculate soil heat conductivity (Yang et al., 2011) but the dry thermal conductivity

AGU Water Resources Research

10.1002/2017WR020635

**Table 1**
*Details of the Key Physics Formulations in the 8 LSMs Used in This Study*

| Physics | Noah28 | Mosaic | VIC403 | SAC | Noah36 | CLSM25 | VIC412L | NoahMP |
|---|---|---|---|---|---|---|---|---|
| Soil hydrology | 4 soil moisture layers (0–10, 10–40, 40–100, 100–200 cm) | 3 soil moisture layers (0–10, 10–40, 40–200 cm) | 3 soil moisture layers (0–10cm; other two layer depths varies spatially) | 6 conceptual soil moisture zones (converted to Noah layers) | 4 soil moisture layers (0–10, 10–40, 40–100, 100–200 cm) | Catchment hydrology with 3 regions: saturated, transpiration and wilting; top soil layer 0–2 cm; root zone layer 0–1 m; depth to bedrock varies spatially | 3 soil moisture layers (0–10 cm; other two layer depths varies spatially) | 4 soil moisture layers (0–10, 10–40, 40–100, 100–200 cm) |
| Canopy interception (capacity; mm) | 0.5 | 0–1.6 | 0.1–1.0 | N/A | 0.5 | 0–1.6 | 0.1–1.0 | 0.5 |
| Vegetation transpiration | Jarvis (1976) | Sellers et al. (1986) | Jarvis (1976) | N/A | Sellers et al. (1986) | Sellers et al. (1986) | Jarvis (1976) | Ball-Berry (Ball et al., 1987) |
| Soil thermodynamics | 4 soil temperature layers (same as soil moisture); heat conduction equation | Force-restore | 3 soil temperature layers (same as soil moisture); heat conduction equation modified | N/A | 4 soil temperature layers (same as soil moisture); heat conduction equation | 6 soil temperature layers; heat conduction equation | 3 soil temperature layers (same as soil moisture); heat conduction equation modified | 4 soil temperature layers (same as soil moisture); heat conduction equation |
| Snowpack physics | 1 snow model layer | 1 snow model layer | 2 snow model layers | 1 snow model layer | 1 snow model layer | 3 snow model layers | 2 snow model layers | 3 snow model layers |
| Runoff generation | Surface and free drainage | Surface and free drainage | Surface and free drainage | Surface and free drainage | Surface and free drainage | Surface and free drainage with shallow groundwater | Surface and free drainage | Surface and baseflow including interaction with groundwater (TOPMODEL) |
| Evapotranspiration | Modified Penman-Monteith | Penman-Monteith | Penman-Monteith | Climatological | Modified Penman-Monteith | Penman-Monteith | Penman-Monteith | Modified Penman-Monteith |

($K_{t,dry}$) and the saturated thermal conductivity ($K_{t,sat}$) are determined using different empirical equations than the VIC model.

All of these LSMs use Richards (Richards, 1931) equation or its variants to describe the soil moisture dynamics as follows:

$$\frac{\partial \Theta}{\partial t} = \frac{\partial}{\partial z}\left(D\frac{\partial \Theta}{\partial z}\right) + \frac{\partial K}{\partial z} + F_{\Theta}, \tag{5}$$

where the hydraulic conductivity $K$ and soil water diffusivity $D$ are functions of soil moisture content $\Theta$, and $F_{\Theta}$ represents the source and sink terms, such as evaporation, transpiration, infiltration, and runoff. Similar to the calculation of soil temperature, the parameterization schemes for $D(\Theta)$ and $K(\Theta)$ and the determination of the source sink terms ($F_{\Theta}$) differ across the LSMs. For example, for evapotranspiration computations, CLSM and VIC use the standard Penman-Monteith formulation whereas Noah models use modified Penman-Monteith approaches for potential evapotranspiration, with added differences in the individual model parameterizations.

The determination of runoff in the LSMs is even more different. For example, VIC uses the variable infiltration curve and Arno baseflow curve to determine surface runoff and baseflow while Noah model adopts the Simple Water Balance (SWB) model to calculate surface runoff and use the gravity drainage from the bottom layer to represent baseflow (Chen & Dudhia, 2001). NoahMP and CLSM use the topographic index concepts of TOPMODEL to describe runoff generation (Koster et al., 2000; Niu et al., 2011). Some models have a free drainage bottom layer, such as VIC, Noah and Mosaic, and some models have a conceptual groundwater component, such as NoahMP and CLSM (Koster et al., 2000; Niu et al., 2011). The vertical discretization of the soil column is also quite different in these models. For example, Noah and NoahMP use a four-layer scheme (0.1, 0.3, 0.6, and 1 m), VIC models use a three-layer scheme with fixed thickness (0.1 m) for the first layer and varying thicknesses for the second and third layers and Mosaic uses a three-layer scheme with a surface layer, a root layer, and a recharge layer, of 0.1, 0.3, and 1.6 m, respectively. CLSM employs a nontraditional approach where the vertical soil moisture profile is determined through deviations from the equilibrium soil moisture profile. The soil moisture in a 2 cm surface layer and a 1 m root zone layer are determined from the bulk moisture variables in the LSM. SAC calculates soil moisture using conceptual zones; the NLDAS team has postprocessed the SAC output to provide soil moisture in the same vertical layers as Noah.

The LSMs use different strategies for surface energy partition of the incoming net radiation into latent, sensible, and ground heat fluxes. In VIC and Noah models, the latent heat flux is determined according to evapotranspiration, which is calculated based on the Penman or Penman-Monteith equations. While in Mosaic, CLSM, and NoahMP, latent heat flux is calculated first and then the evapotranspiration is diagnosed. The energy balance calculation is the same in Mosaic and CLSM (Koster et al., 2000), where sensible heat and latent heat fluxes are calculated using a resistance network analog. The version of SAC used in NLDAS-2 does not calculate latent and sensible heat fluxes. Different numeric solutions are applied to determine skin temperature estimates, which also impacts the output energy fluxes of these models.

Snow pack calculation in the LSMs uses similar physics schemes with different vertical discretization of snow columns and different parameterizations for properties such as snow heat capacity and snow heat conductivity. VIC and Noah models use a single layer snow scheme, CLSM uses a three-layer scheme, and NoahMP uses a multilayer discretization scheme. Compared to Noah28, Noah36 includes snow physics updates (University of Arizona snow physics option; Wang et al., 2010) and updates for snow albedo, surface roughness and surface exchange coefficient formulations (Barlage et al., 2010).

In addition to the physics and parameterization differences, there are differences in the land surface parameters employed in the LSMs. All model integrations use the global vegetation classification data set (with 13 classes) produced by the University of Maryland (Hansen et al., 2000) as the landcover data and the State Soil Geographic Database (STATSGO) soils database (Miller & White, 1998) to specify soil texture types in the model configuration. The VIC models use leaf area index (LAI) derived from AVHRR measurements whereas the Noah models use Green Vegetation Fraction (GVF) derived from MODIS data products. The NoahMP configuration used in this article employs a dynamic phenology model where LAI is a model prognostic variable. A set of recalibrated soil hydrology parameters is used in the VIC412L simulation, which also contributes to the differences between the results of VIC403 and VIC412L.

### 2.2. Variables

Six variables (latent heat (Qle) and sensible heat (Qh) fluxes, total runoff (Q), root zone moisture content (RZMC), snow water equivalent (SWE), and terrestrial water storage (TWS) that represent some of the key components of the terrestrial water and energy budget from these eight LSMs are employed in the similarity assessments. The latent and sensible heat fluxes describe the surface energy partition of the available net radiation. Correspondingly, the surface water balance is described by the partition of the incoming precipitation into evapotranspiration, runoff and the change in storage within the soil. The latent heat flux is directly linked to the evapotranspiration component of the surface water budget, as it represents the energy transfer term associated with evaporation and vegetation transpiration. The total runoff employed in the analysis is the sum of the surface and subsurface runoff. SWE, RZMC, and TWS variables represent the key storage terms of the water balance. RZMC is defined for this study as the total soil moisture of top 1 m soil column, determined in each LSM as a suitably weighted average over the model layers. TWS represents the vertically integrated measure of the surface and subsurface water and is computed as the sum of soil moisture, snow water equivalent, canopy water storage, and groundwater. Note that only CLSM and Noah-MP simulates subsurface groundwater reservoirs among the eight LSMs. In all other models, the terrestrial water storage is essentially represented by the components of canopy interception, snow water equivalent, and soil moisture. In the analysis below, we employ daily and monthly averaged estimates of these six variables.

### 2.3. Domain and Time Period

The domain configuration in this study is the one used for NLDAS-2, which consists of a 0.125° lat by 0.125° lon grid that extends from 25°N to 53°N and 125°W to 67°W. The similarity assessments are conducted using model outputs over an eleven year time period, from January 2002 to December 2012. All the models are forced with the NLDAS-2 forcing data (Xia et al., 2012b) and run at 15 min time intervals except VIC403 and VIC412L which use a 1 h time step. All models were spun-up by initializing with that model's equilibrium state at the start of the NLDAS-2 forcing in January 1979, and run for 23 years before the start of the analysis period of January 2002. The NASA Land Information System (LIS; Kumar et al., 2006; Peters-Lidard et al., 2007) is used to conduct the model simulations with the four new LSMs.

### 2.4. Reference Data Products

To develop assessments of accuracy, the model estimates of Qle, Qh, Q, RZMC, and TWS are compared against a number of available reference data sets for these variables. The model Qle and Qh outputs are compared against in situ flux measurements from 76 AmeriFlux network stations. These sites are chosen from the AmeriFlux Level 3 data, which is a quality controlled version of the raw measurements. The flux measurements available at 30 min intervals are aggregated to a daily time scale for comparison against the LSM estimates. Soil profile measurements from the USDA Soil Climate Analysis Network (SCAN; Schaefer et al., 2007) are used as the reference data for evaluating RZMC. The SCAN network stations provide hourly soil moisture measurements at depths of 5, 10, 20, 50, and 100 cm wherever possible. For evaluating total runoff estimates from the LSMs, the gridded Daily streamflow data obtained from the United States Geological Survey (USGS; http://nwis.waterdata.usgs.gov/nwis) over 572 small, unregulated basins is used to evaluate the total runoff estimates from the LSMs. These basins were also part of the model evaluations used in the NLDAS-2 project (Kumar et al., 2014, 2016; Xia et al., 2012c) and are a subset of the Model Parameter Estimation Experiment (MOPEX) study basins. The TWS estimates from the LSMs are evaluated against TWS anomalies from the Gravity Recovery and Climate Experiment (GRACE; Tapley et al., 2004) mission. In this study, we employ the monthly gridded (available on 1° horizontal resolution grids during January 2003 to January 2013) Tellus GRACE TWS anomaly products (Landerer & Swenson, 2012). This product is bases on the Release-05 (RL05) spherical harmonics fields produced by the University of Texas Center for Space Research (CSR). The scaling coefficients provided with the data for restoring some of the signal loss due to filtering and truncation during the TWS derivation are also applied in the evaluations presented here. Note that we exclude an accuracy evaluation of modeled SWE estimates due to the limited availability of SWE reference data sets. Though SWE measurements are available from the Natural Resources Conservation Service (NRCS) Snow Telemetry (SNOTEL) network, the SNOTEL stations tend to be located in high-elevation mountain watersheds. As a result, the SNOTEL SWE measurements are more representative of the SWE extremes rather than the mean. The SNOTEL network is also limited in its coverage to the Western U.S.

### 2.5. Assessment Methods

We apply a latent variable model employing a confirmatory factor analysis (CFA; Christensen & Sain (2012); Hattie & Fraser, 1988) to quantify unmeasured sources of similarity and variability in the model predictions. A latent variable model is a statistical approach that relates observed variables to latent or unobserved variables and can be represented through a linear regression model as follows:

$$x_k = \mu + \lambda_k f + e_k, \tag{6}$$

where $x_k$ is the modeled estimate for a given variable from model $k$, $\mu$ is a constant to all models, $e_k$ is an independent term specifically related to model $k$, and $f$ is the standardized common factor (latent variable) across all models. The regression coefficient $\lambda_k$ is called the factor loading for model $k$ that represents how strongly the observed variable is associated with the common factor. For each variable, $f$ is unique and is assumed to be normally distributed with zero mean, uncorrelated with $\lambda_k$ and a unique variance (Cai, 2012). The factor loading $\lambda_k$ ranges from $-1$ to 1. Factor loadings close to $-1$ or 1 indicate that the common factor strongly affects the variable. On the other hand, factor loadings close to zero indicate that the common factor has a weak effect on the variable. In other words, if a model has a factor loading close to 1, it indicates that the model output is very close to the common factor of the ensemble. If there are two models with factor loadings close to 1, it means that the two model outputs are similar to each other. However, if a model has a negative factor loading, it indicates that the model output has a trend opposite to the common factor of the output ensemble (the model output is dissimilar to others). In the following sections, we employ the latent variable model at different time scales to examine the similarity/dissimilarity of LSMs.

To quantify the accuracy of model estimates, the anomaly RMSE metric is used for Qle and Qh whereas the anomaly correlation (anomaly R) measure is used for RZMC, Q, and TWS. Since the gridded runoff is compared against the routed streamflow, a direct comparison metric such as RMSE is less meaningful. Similarly, consistent with prior studies (Kumar et al., 2012, 2014), we employ the anomaly R metric to evaluate RZMC given the significant differences in the climatologies of the soil moisture fields from each model and the reference data sets. The anomalies are computed by subtracting the monthly mean climatology of each variable from the corresponding daily average raw data. In using the anomaly based metrics, the skill of the mean seasonal cycle is excluded in the accuracy quantifications.

## 3. Results

### 3.1. Intermodel Similarity Assessment

Figure 1 shows a comparison of the average anomaly correlation (R) among the eight LSMs for the daily averaged Qle, Qh, Q, RZMC, SWE, and TWS. These plots are generated by first computing an $8 \times 8$ matrix of anomaly R values for different pairs of model comparisons (Noah28 versus Mosaic, Noah28 versus VIC, Noah28 versus Noah36, and so on), for each variable. Each panel in Figure 1 shows the average of the off-diagonal elements of the $8 \times 8$ matrix, representing the intermodel correlations. These average anomaly R maps essentially represent a first order estimate of where the constituent models agree with each other and by how much. For example, for Qle, there are both areas of strong agreements (Central and Southwest U.S.) and disagreements (Southeast, mountainous regions in the Western U.S., urban areas), whereas in the Q comparison, the models show large disagreements in the Western U.S. and better agreements in the Eastern U.S. Figure 1 also shows the spatial distribution of the average anomaly R for each variable. Overall, the model estimates are most dissimilar in their runoff estimates and most similar for TWS, RZMC and SWE estimates. The level of agreements and disagreements is moderate for the energy budget terms of Qle and Qh.

To quantify the spatial variability of the intermodel similarity, the Köppen-Geiger climate classification is used to stratify the average anomaly R maps into 10 different climatic zones. Figure 2 (top) shows the updated Köppen-Geiger climate classification for the NLDAS-2 domain, which is based on the 0.5° global climate map, generated using 50 years (1951–2001) of climate observations (available online at http://koeppen-geiger.vu-wien.ac.at/present.htm). There are five main climate zone groups in the Köppen climate classification scheme. They are A, equatorial; B, arid; C, warm temperate; D, snow; E, polar. Within each of these primary climate zones, further categorization based on precipitation and temperature regimes are included. Because of very small coverage over the NLDAS-2 domain, a number of these climate zones from the original data have been merged into other categories, the details of which are shown in Table 2.

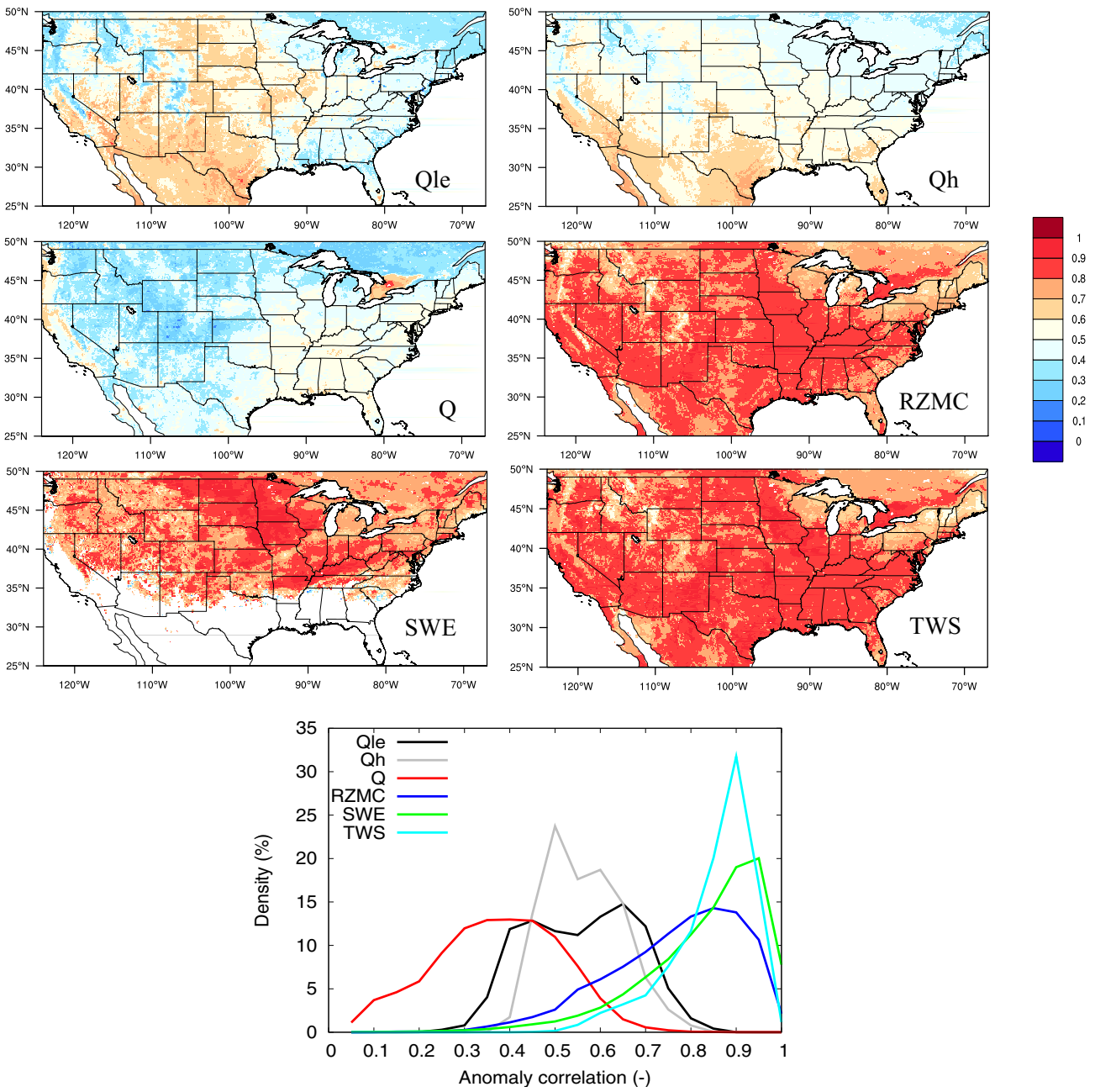


**Figure 1.** Average anomaly correlation among the eight LSMs for latent heat flux (Qle), sensible heat flux (Qh), total runoff (Q), root zone soil moisture (RZMC), and snow water equivalent (SWE). The bottom right figure shows the distribution of the average anomaly R across the domain for each variable. The anomaly R values are computed using daily averaged time series for each variable.

Figure 2 (bottom) shows the variability of average anomaly R values across Köppen-Geiger climate classification zones Koppen (2011). For Q, the average anomaly R values are generally larger (i.e., the LSMs agree with each other more) in the warm temperate zones (Cfa, Cfb, and Csb) relative to the arid zones (BSh, BSk, BWh, and BWk). The warm temperate and humid zones (Cfa and Cfb) receive more precipitation compared to the arid zones and the runoff generation is generally dependent on the soil moisture alone. In the cold regions (Dfa, Dfb, and Dfc), more variability in the average anomaly R values is seen with a gradual decrease in the agreement among models in colder regions. The stratification of the average R values for fluxes (Qle and Qh) shows similar behavior across the climate zones with greater agreement among the models in the arid zones and greater disagreements among the models in the cold regions. Compared to the dry regions,

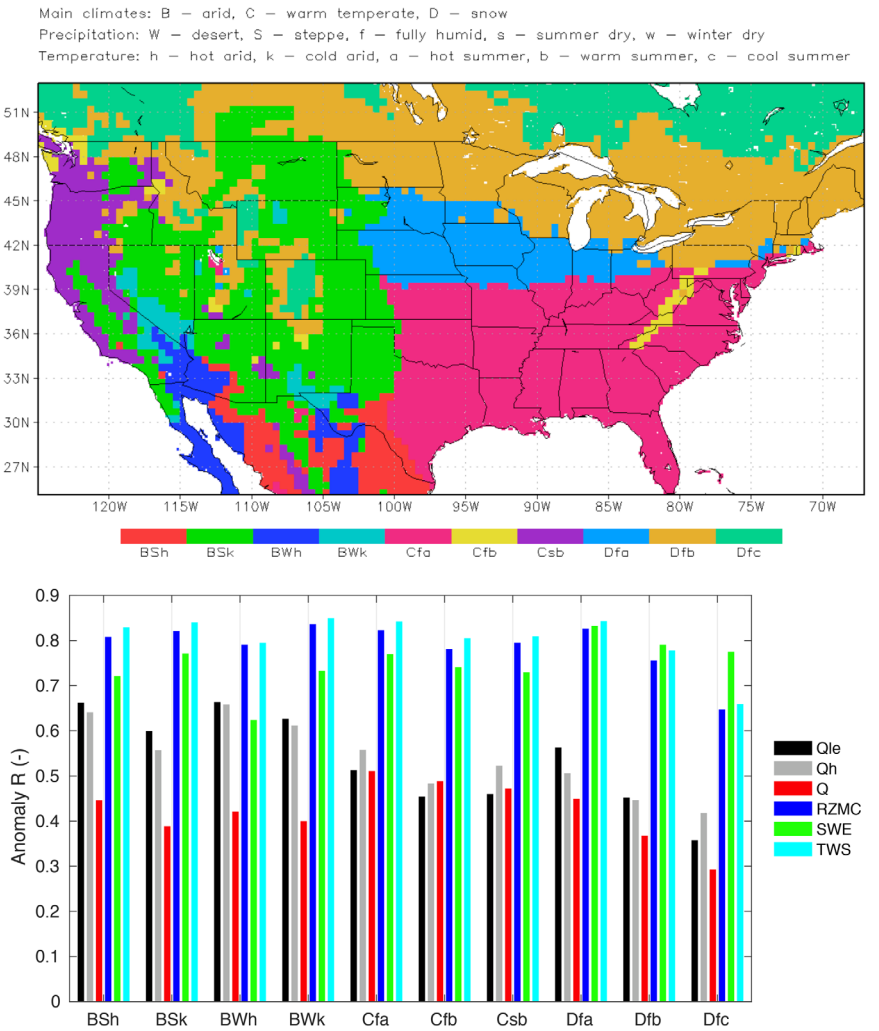


**Figure 2.** (top) Simplified Köppen-Geiger climate classification zones for the NLDAS-2 domain and (bottom) average anomaly R values for five variables stratified using the Köppen-Geiger climate classification zones over the NLDAS-2 domain.

vegetation and cold season processes are more impactful in the surface energy partition in the warm temperate and cold regions, leading to larger dissimilarity among the models. For SWE, the models show high levels of similarity, especially in the cold regions. Highest average anomaly R values for SWE are seen in the

**Table 2**
*Remapped Köppen-Geiger Climate Classification Zones in the NLDAS2 Domain*

| Original classification | | Updated classification |
|---|---|---|
| Equitorial climate (Af, Am, As, Aw) | → | Fully humid warm temperate with hot summer (Cfa) |
| Snow climate with dry and cool summer (Dsc) | → | fully humid snow climate with cold summer (Dfc) |
| Polar tundra (ET) | | |
| Snow climate with dry and warm summer (Dsb) | → | Fully humid snow climate with warm summer (Dfb) |
| Snow climate with dry winter and warm summer (Dwb) | | |
| Snow climate with dry winter and hot summer (Dwa) | → | Fully humid snow climate with hot summer (Dfa) |
| Warm temperate climate with dry winter and warm summer (Cwb) | → | Bsk (North of 30°N) |
| | | Bsh (South of 30°S) |

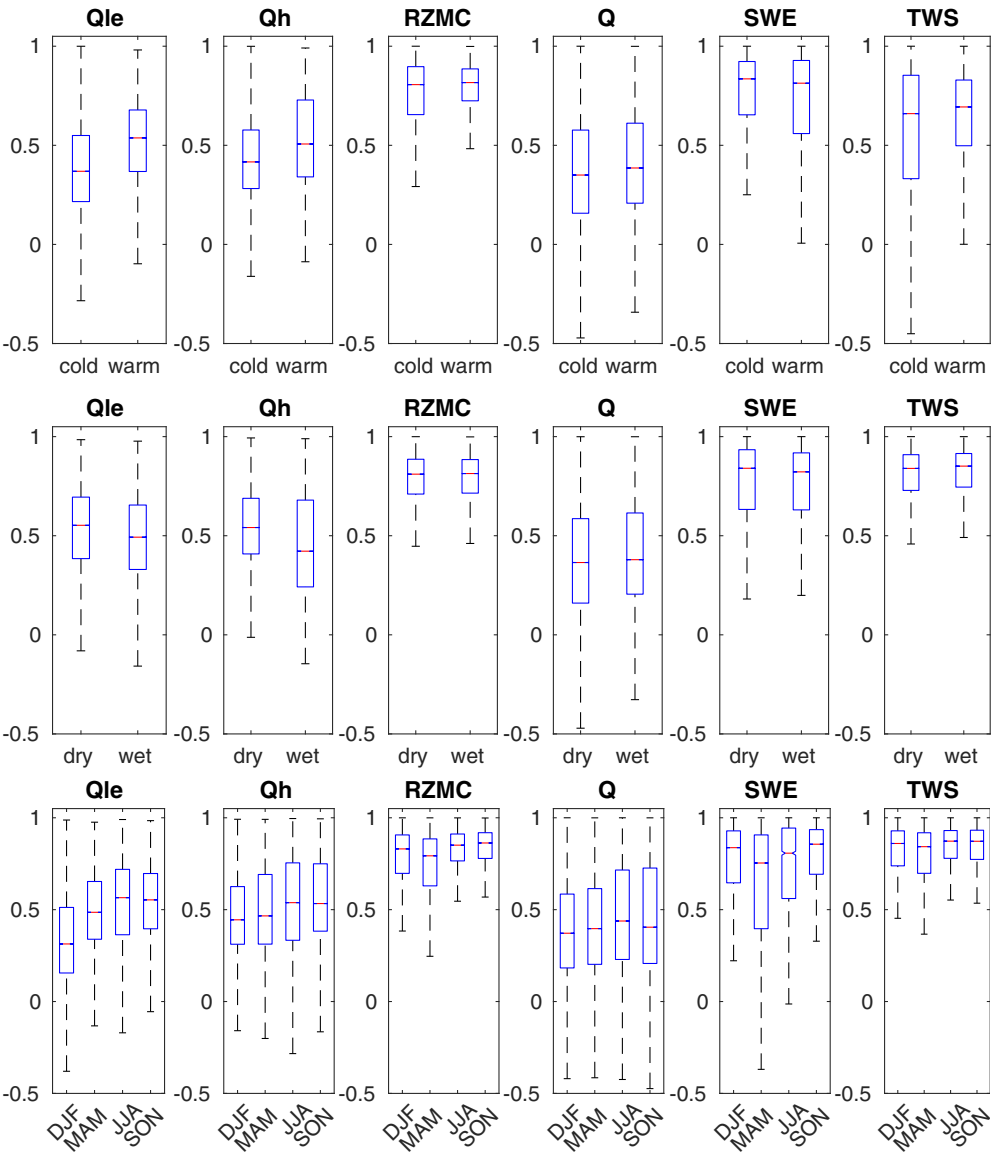


**Figure 3.** Temporal stratification of the intermodel average anomaly R values. (top) The stratification for cold and warm days, (middle) the stratification for dry and wet days, and (bottom) the stratification for four (DJF, MAM, JJA, and SON) seasons.

Dfa and Dfb zones, cold regions dominated by warm summers. Figure 2 shows high levels of agreement among the LSMs for RZMC estimates, particularly in the arid and warm zones. The level of similarity reduces in the cold regions, especially in the regions with prolonged winter (Dfc). Similar to RZMC, the average anomaly R values for TWS are also high across different climate zones, as RZMC is the primary contributor of TWS variability in most LSMs.

Figure 3 shows three different temporal stratifications of the average anomaly R calculations. The top row indicates boxplots of the distributions of anomaly R values for warm and cold days, where a warm day is defined to be when the daily averaged NLDAS-2 forcing 2 m air temperature is greater than 273.15 K. The models show increased agreement for Qle, Qh, Q, RZMC, and TWS during warm time periods whereas for a cold season process such as SWE, greater agreement among the models is observed for cold days. The middle row of Figure 3 shows the distribution of the average anomaly R values stratified for wet (when total precipitation is $>0$) and dry days. The models show comparable levels of agreement in Qle, Q, SWE, RZMC, and TWS, for both dry and wet days whereas the Qh estimates among the models show more dissimilarity

over wet days. Finally, the bottom row shows the stratification of the average anomaly R values for four seasons. Similar to the trends seen in the top row, estimates of Qle and Qh fluxes among the models are most dissimilar during the winter time periods. For SWE, RZMC, and TWS, the model estimates differ most during the spring melt time periods (MAM). In general, the distribution of TWS is similar to that of RZMC in these temporal stratifications, confirming the prominent contribution of the soil moisture term in the LSM TWS estimates. The influence of SWE to the TWS calculation can be noted in the warm-cold stratification where the TWS distribution is similar to RZMC, but with a noticeably larger range.

### 3.2. Factor Analysis

Comparison of the average anomaly R maps presented in the previous section provides an overall measure of the similarity among the LSMs for different variables. In this section, a latent variable model is employed to isolate the level of similarity of each LSM to the common factor across all models. Specifically, the factor loadings ($\lambda_k$) quantify the level of closeness of each individual model estimate to the common factor across all LSMs. High factor loading values indicate a close association of the individual model estimate to the common factor and values closer to zero indicate a weak association with the common factor.

Figure 4 shows the spatial maps of the factor loadings for Qle computed based on the daily anomalies of the LSMs. Note again that SAC does not calculate Qle (or Qh) and SAC is not included in the calculation of the factor loadings for Qle and Qh. Overall, Noah36, CLSM25, and NoahMP have high factor loadings among the LSMs, whereas Noah28 and Mosaic LSMs show large spatial variability in the range of values. The factor loading patterns are relatively uniform in the VIC403 LSM whereas the new version of the model (VIC412L) shows larger values, suggesting that the VIC412L model estimates are closer to the common factor. Compared to the Southeast U.S., larger factor loadings are generally observed over the Southwest U.S. for most LSMs, indicating that the individual model estimates are closer to the common factor in water limited domains, whereas they differ more in energy limited domains, with VIC403 as the notable exception that does not show such a contrast. The factor loadings also show more variability over areas where cold season processes are prominent.

A similar comparison of the factor loadings for Qh from individual models are shown in Figure 5. Similar to the trends in Figure 4, the newer LSMs, especially Noah36, CLSM25, and NoahMP show larger factor loadings. For Noah36 and NoahMP, there is little spatial variability in the factor loadings. Generally, the models tend to differ more from the common factor over the water limited domains (Southwest) compared to the energy limited domains (Southeast). The notable exception to this trend is VIC412L, which shows a high degree of agreement with the common factor in the Southeast and Southwest (especially over California and Arizona) U.S. based on the average factor loading values. Overall, VIC403 is the most dissimilar model in the surface flux comparisons. Figures 4 and 5 also indicate that there are significant differences in the spatial distribution of the factor loadings for Qle and Qh between the NLDAS-2 operational models and the newer LSMs. The areas of similarity and dissimilarity are distinct in factor loading maps of the NLDAS-2 operational models, with the relative agreement with the common factor stronger in the dry and warm climates and weaker in the colder climate regimes. Such contrasts in the factor loadings are less distinct in the newer LSMs, when stratified based on the K-G climate zones (not shown).

Figure 6 shows the factor loading spatial maps of the LSMs for Q. Overall, the LSMs fall into two distinct groups with the NLDAS-2 operational models showing weak associations to the common factor and the new LSMs showing closer alignments. Across the NLDAS-2 operational models, larger factor loadings are obtained over the Southeast U.S. and over the West Coast, whereas smaller factor loadings are seen over the Central U.S. In contrast, the newer LSMs show strong agreement with the common factor over the Central U.S., with smaller factor loading values over the Southeast and West Coast. Based on the average factor loading values, Noah28 and Mosaic LSMs show largest departures from the common factor whereas VIC412L shows the strongest agreement. When the factor loadings for each model are stratified based on K-G zones, the NLDAS-2 operational models show stronger agreement in the warm and temperate zones (Cfa, Cfb, and Csb) compared to the dry and cold regimes (not shown). The new LSMs, however, show less contrast between different climatic regimes with the generally high degree of agreement with the common factor.

For RZMC, Figure 7 shows the spatial maps of factor loadings for the LSMs. Relative to the factor loading comparisons of Qle, Qh, and Q, all models generally show strong agreement with the common factor

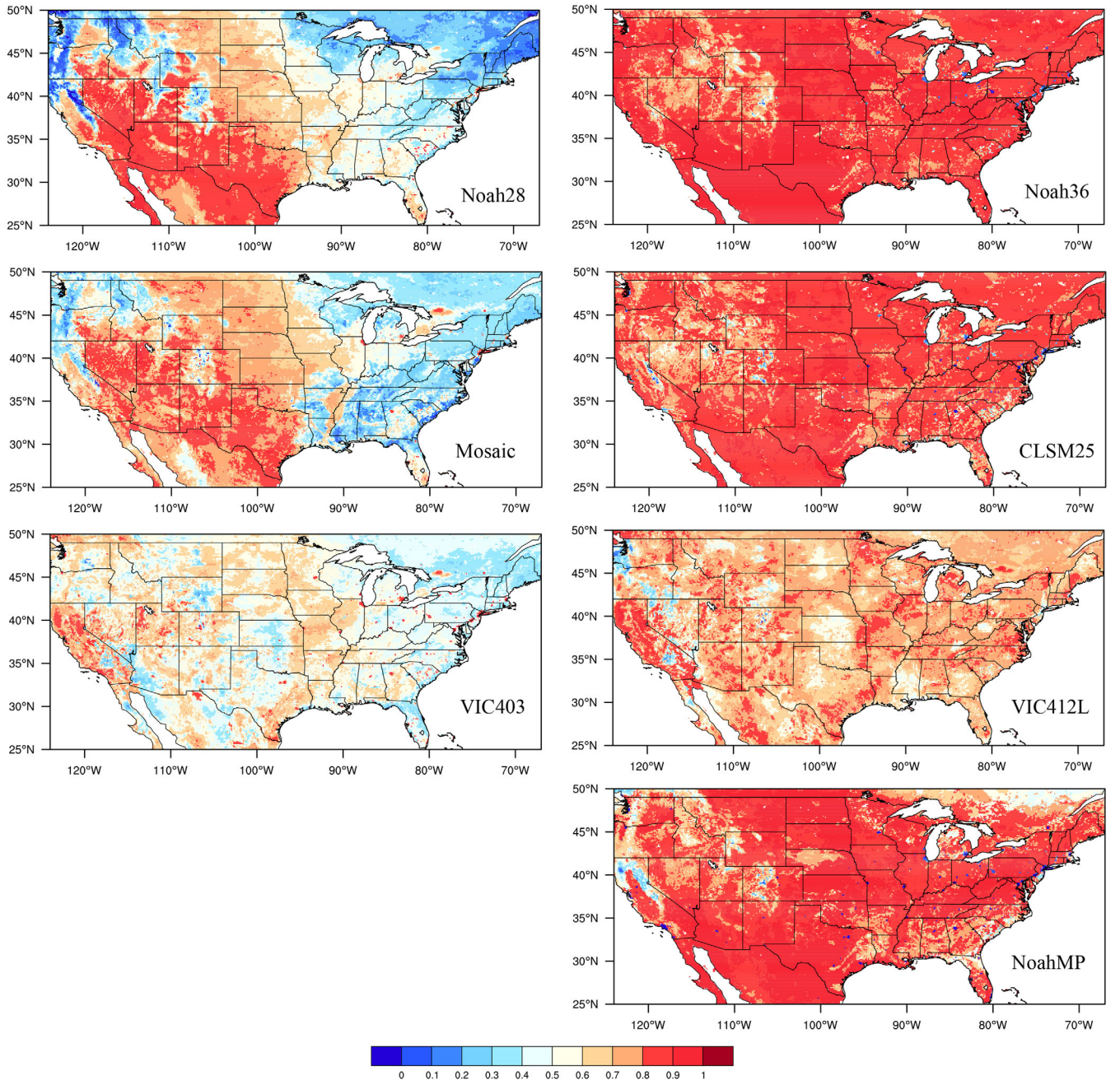


**Figure 4.** Factor loadings of the land surface models for latent heat flux. Values near 1 indicate a strong association of the particular LSM to the common factor shared by the LSMs and values near zero indicate weaker association of that model with the common factor.

(VIC412L with the largest average factor loading and CLSM25 the lowest). The contrast between the NLDAS-2 operational LSMs and the newer LSMs is less obvious in this comparison. The most significant differences in the factor loadings are observed in regions where cold season processes play an important role. The stratification of the factor loadings to the K-G zones confirms these trends (not shown).

Figure 8 shows the factor loading comparison across the LSMs for SWE. Similar to the case for root zone soil moisture, the models show strong agreement with the common factor. Over the dry and warm zones, the agreement among the LSMs is strong. Over the cold regions, CLSM25 shows the largest departures from the common factor. The temporal variations of CLSM25 SWE estimates are weakly relevant to the common factor of the SWE ensemble. Overall, Noah36 and CLSM25 show the closest alignment and dissimilarity,

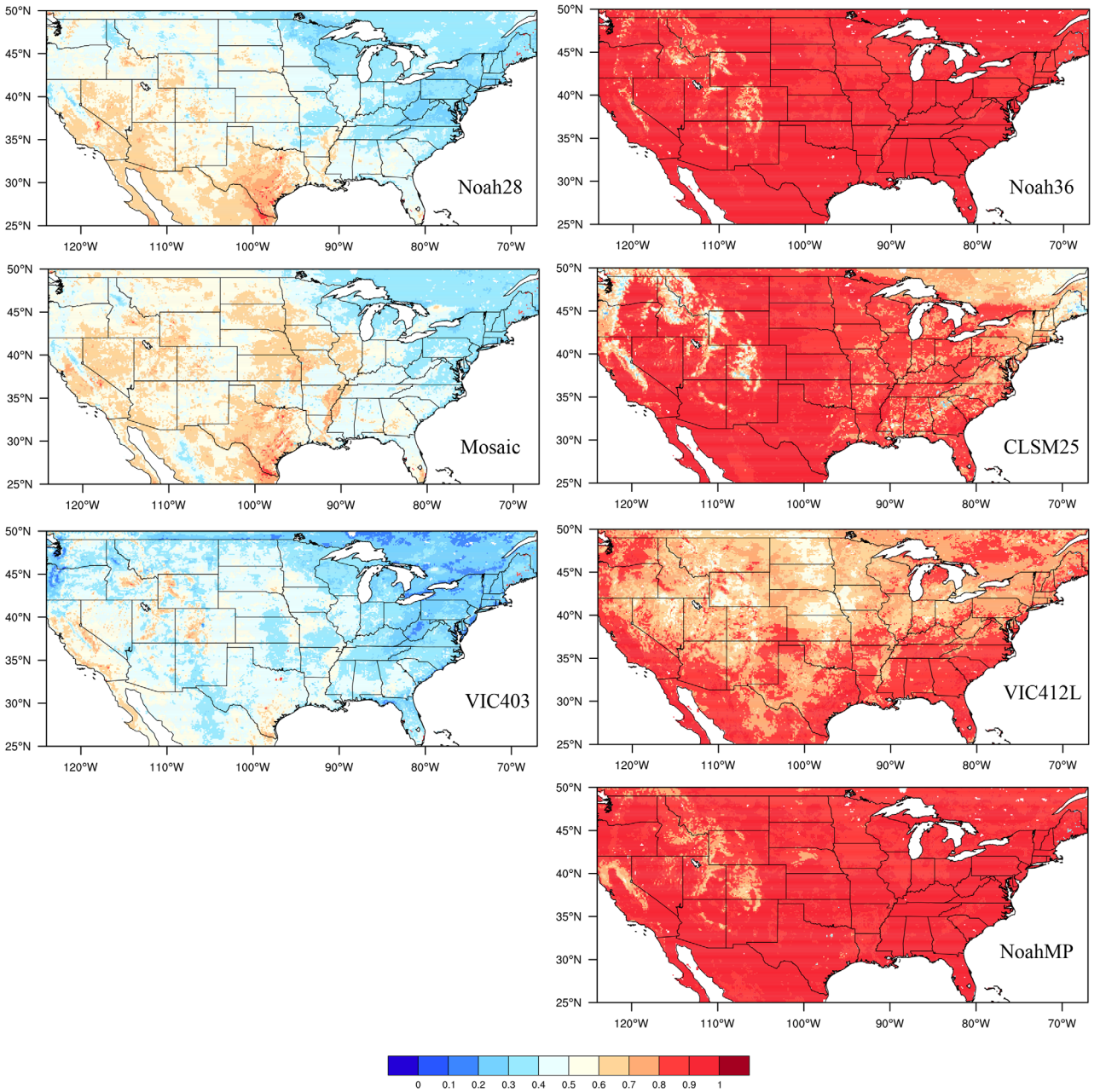


**Figure 5.** Similar to Figure 4, but for sensible heat flux.

respectively, with the common factor. Finally, Figure 9 shows the factor loadings for TWS. All models show very strong agreement with the common factor. This is not surprising given that soil moisture and SWE components are the primary components of TWS, which also show high factor loading values. These trends are also consistent with prior studies (Xia et al., 2017) that have established that soil moisture is the primary contributor to TWS in most traditional LSMs.

A summary of the analysis of factor loadings are presented in Figure 10, which shows a spatial map of the LSM that is most similar and dissimilar to the common factor, for each variable. For surface fluxes, the strongest association with the common factor is shown by NoahMP over the Eastern U.S. and Highplains, CLSM25 over the Southeast and parts of Western U.S., whereas Noah.3.6 is dominant at other regions. On the other hand, the weakest association with the common factor (for Qle and Qh) is shown by the NLDAS-2 operational models, mainly VIC403 and Noah.2.8. The strong association map for Q shows large spatial

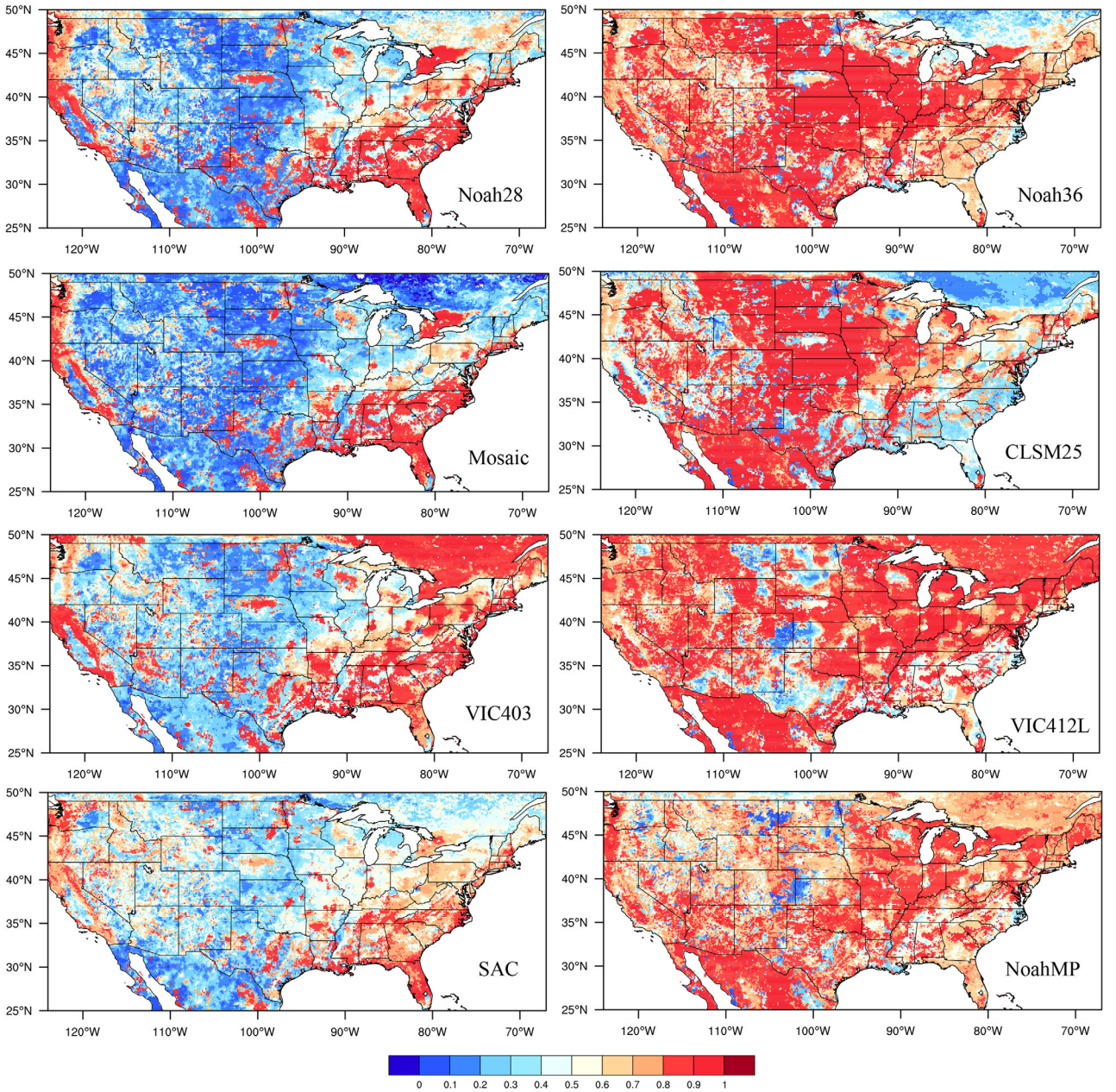


**Figure 6.** Similar to Figure 4, but for total runoff.

variability and the weak association map for Q is dominated by Mosaic and CLSM25. Interestingly, CLSM25 features prominently in both the strong association map (parts of Central and Western U.S.) and the weak association map (Southeast U.S., West Coast). The strong association map for RZMC encompasses most models, whereas the weak association map is dominated by NoahMP and CLSM25. In the map for SWE in Figure 10, NoahMP and Mosaic show the weakest alignment with the common factor over the Highplains and CLSM25 over other regions. The strong association map of SWE is dominated by Noah36. Reflective of the trends in RZMC and SWE, the TWS strong and weak association maps encompass most LSMs. The comparison of the factor loadings for different variables presented above indicate that the level of association of constituent models to the common factor varies significantly based on the LSM, climate regime and the variable, though the driving meteorology is the same. It also indicates that the contribution of a particular LSM to the overall ensemble also depends on the variable under consideration.

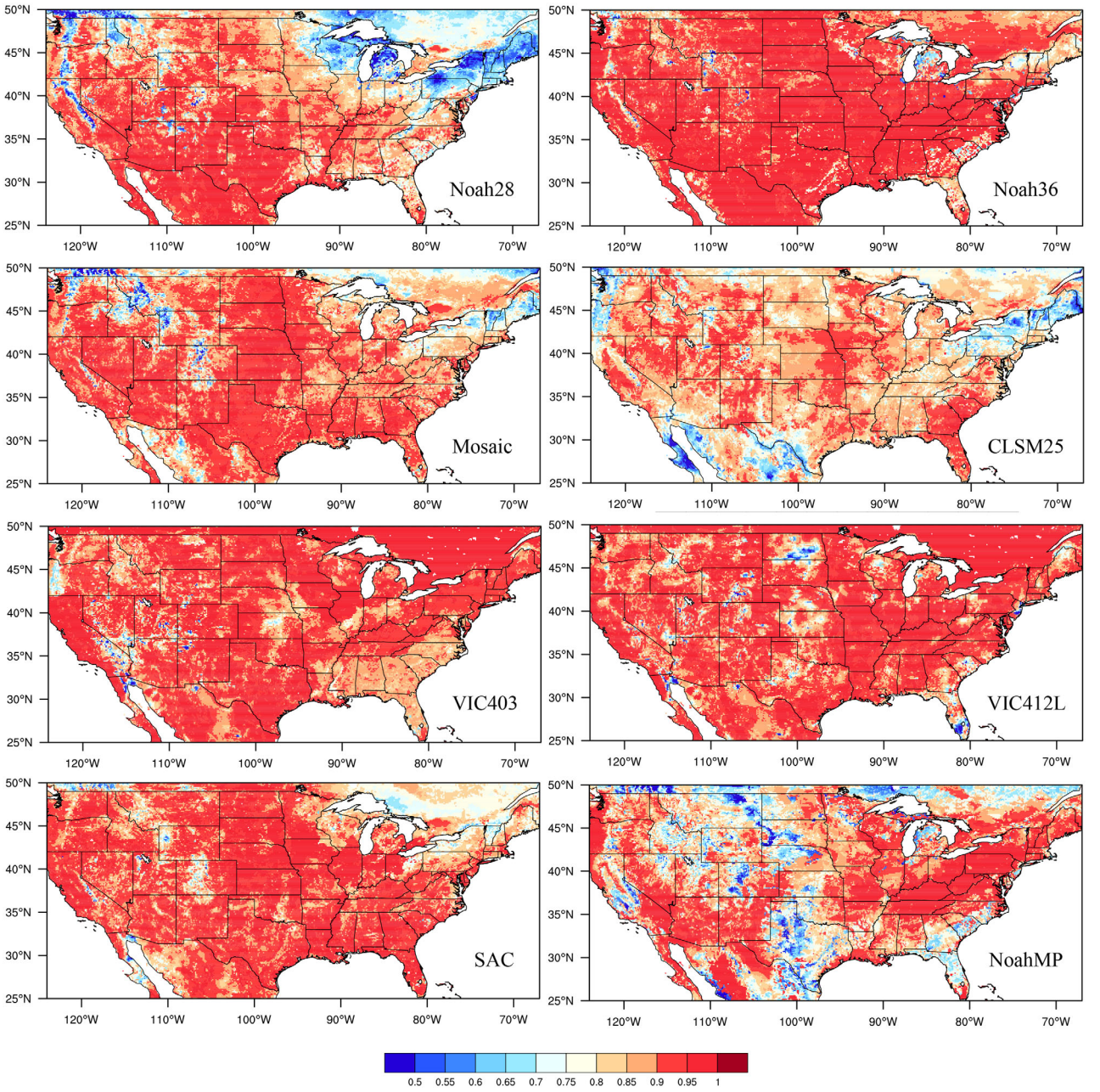


**Figure 7.** Similar to Figure 4, but for root zone soil moisture.

## 4. Similarity Assessments at Different Time Scales

The NLDAS-2 outputs are used for a variety of modeling and research applications ranging from watershed and water quality management to drought and flood monitoring. As the relevant time scales of these processes vary significantly, it is important to assess the usefulness of the ensemble at different time scales. An evaluation of the model similarity at monthly time scale is presented in this section.

Figure 11 shows a comparison of the distribution of the factor loadings for each model computed based on the daily and monthly anomaly correlations. In most cases, the density of grid points with weak association with the common factor reduces and the factor loadings distribution shifts to the right when the time scale of computations is switched from daily to monthly scale, indicating that the LSM estimates get closer to the common factor at the monthly scale. A notable exception is the comparison for CLSM25, where the

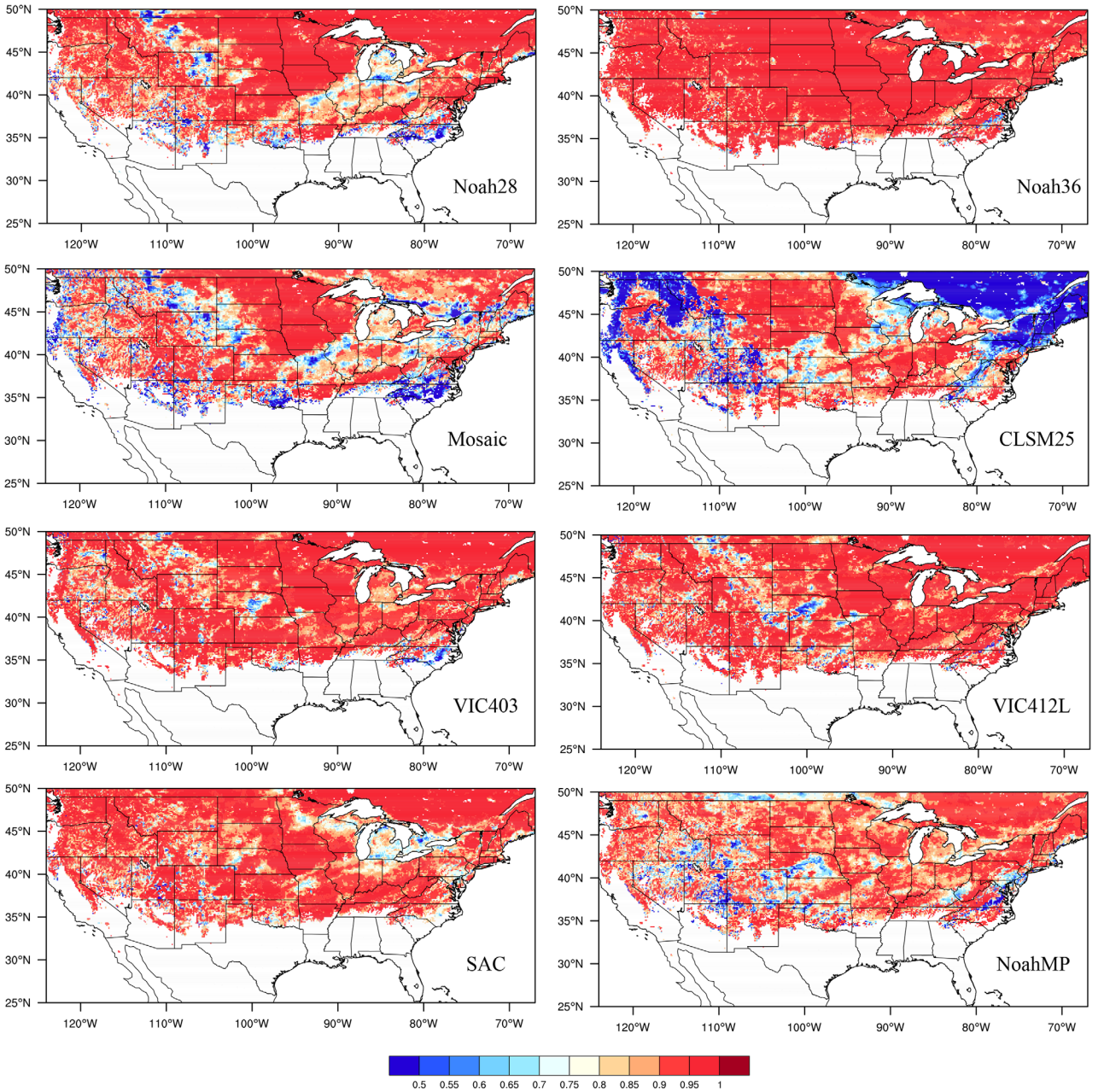


**Figure 8.** Similar to Figure 4, but for snow water equivalent.

distribution of factor loadings for Qh, Q, and RZMC are comparable at both time scales, whereas Qle and SWE estimates are closer to the common factor at the monthly time scale. The analysis suggests that the estimates from the NLDAS-2 operational models and the newer LSMs are more similar at the monthly scale. In other words, the individual contribution of the constituent models to the ensemble is significantly lower at the monthly time scale.

## 5. Evaluation of Model Similarity in Relation to Accuracy

The factor analysis presented in the previous sections is a quantitative way to assess the relative contribution of the constituent models to the ensemble. The utility assessment of a model to the ensemble, however, must also consider the accuracy of the estimates. Ideally, the ensemble must be composed of models

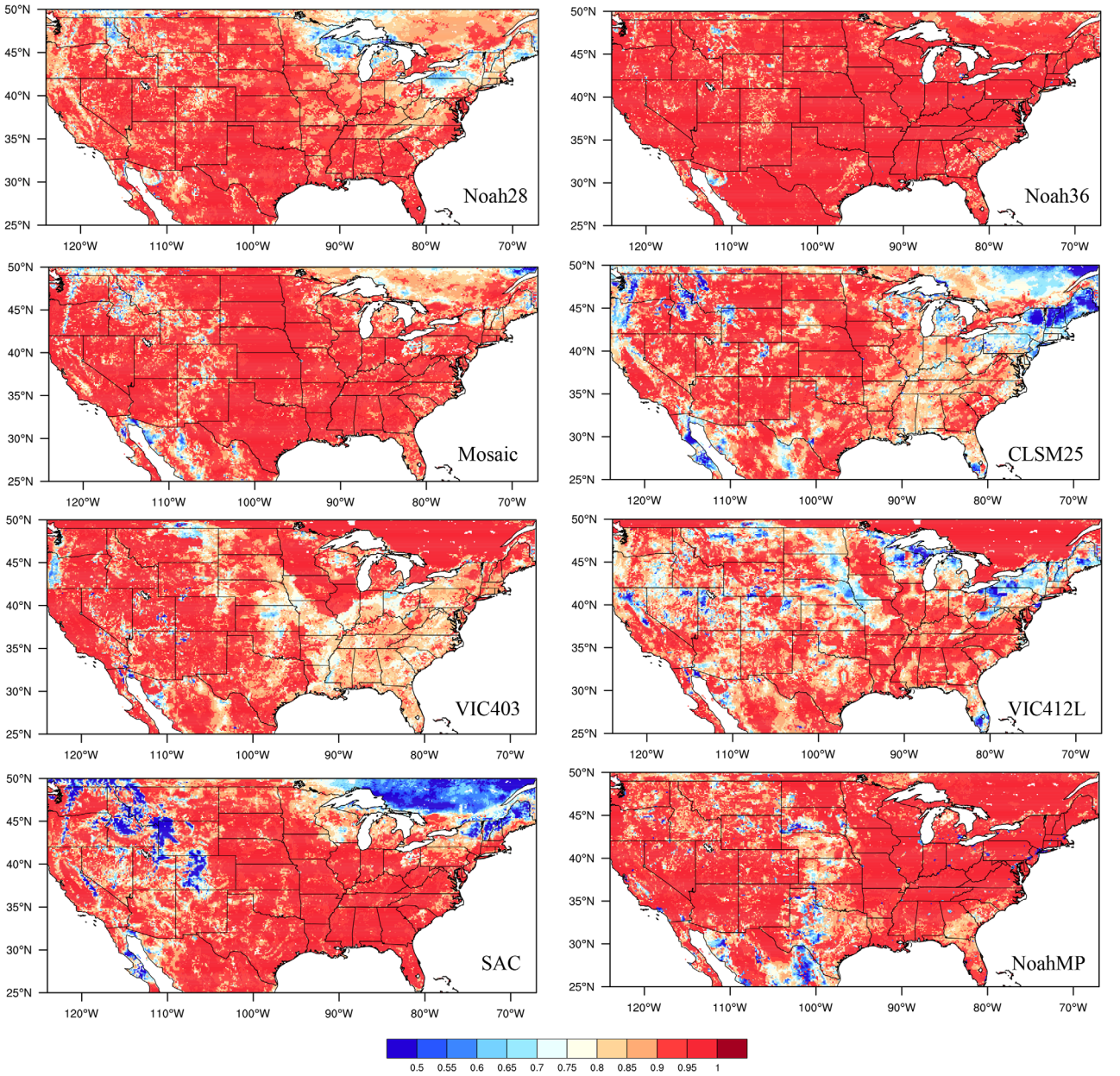


**Figure 9.** Similar to Figure 4, but for terrestrial water storage.

that encompass different regimes in the similarity space, but with high accuracy. In this section, we present a trade-off analysis that compares the models in relation to their accuracy and similarity.

The model performance is quantified as a function of two factors: (1) the model similarity and (2) the model accuracy for these three variables. Figure 12 shows a heatmap/density of grid points as a function of these two factors for latent heat flux. In order to compare different models, the factor loadings values of each LSM are normalized based on their max/min values and are used as analogs for normalized similarity ($NS^i$), with values ranging from 0 to 1 (equation (7)), where $k$ represents a given LSM and $\lambda^i_{max}$ and $\lambda^i_{min}$ are the maximum and minimum factor loading values across all LSMs, at a given grid point $i$. Similarly, the anomaly R (for RZMC and TWS; equation (8)) and anomaly RMSE (for Qle, Qh, and Q; equation (9)) values are normalized to generate a normalized accuracy ($NA^i_{aRMSE}$ and $NA^i_{aR}$) measure with a range of 0–1. Note that higher the factor loading, the normalized similarity values are closer to 1 and the lower the factor loading value, the $NS$ values get closer to 0. Similarly, normalized accuracy values closer to 0 and 1 indicate estimates with

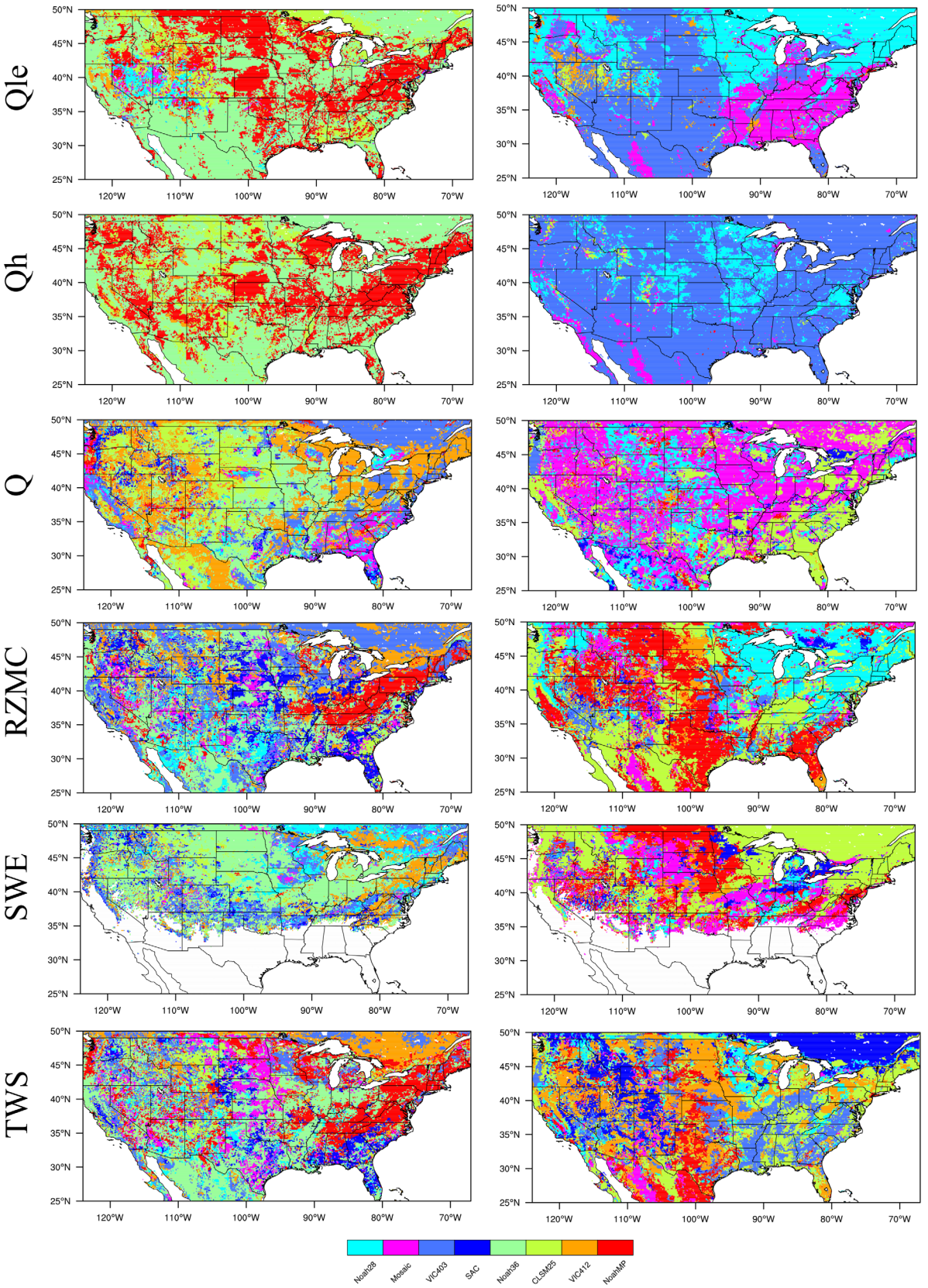


**Figure 10.** Map of LSM most (left column) similar and (right column) dissimilar to the common factor, for Qle, Qh, Q, RZMC, SWE, and TWS.

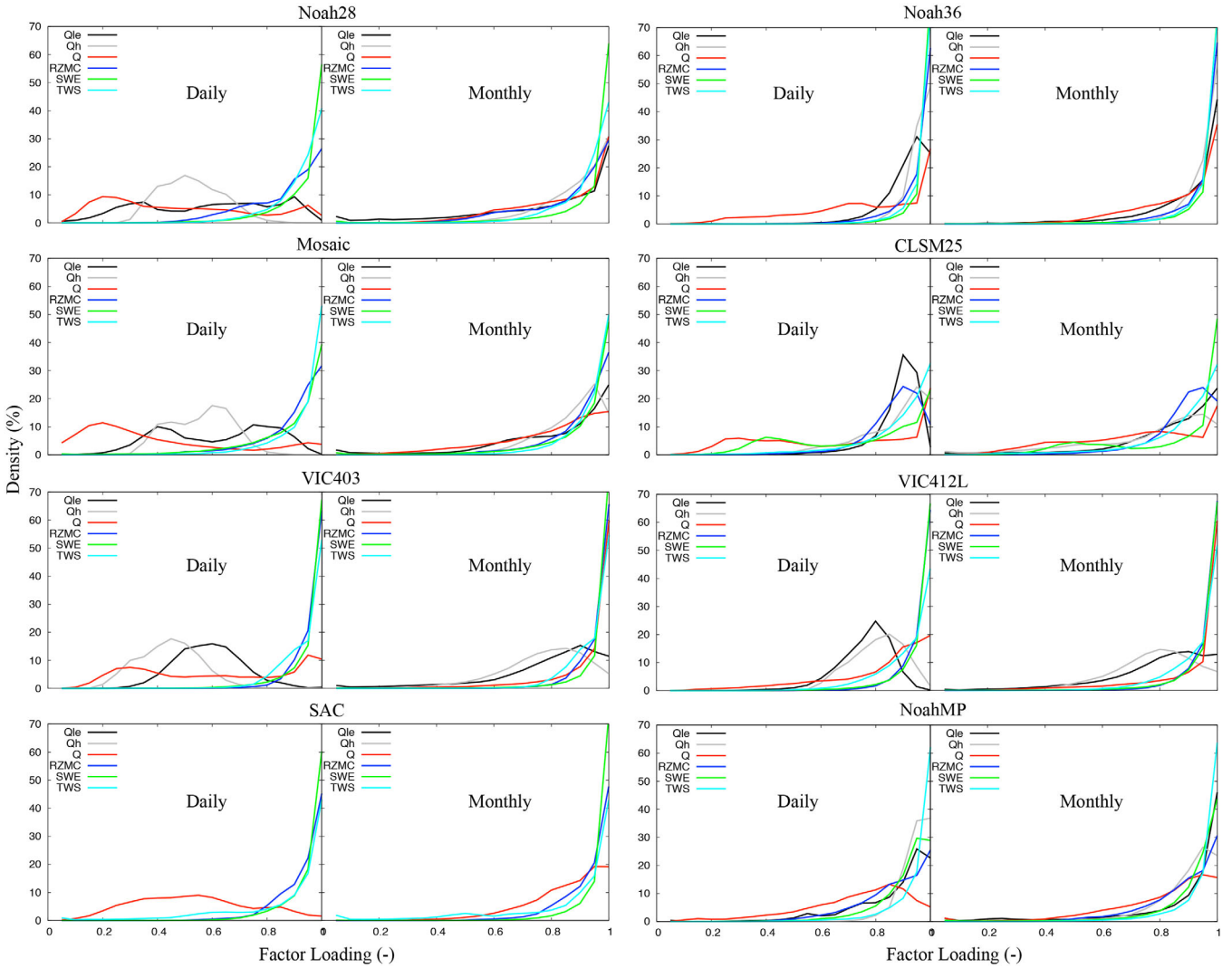


**Figure 11.** Distribution of the factor loadings for each LSM computed based on daily and monthly anomaly R values.

low and high accuracy, respectively. In the trade-off maps based on these normalized measures, the top right corner represents areas with high accuracy and high similarity and the bottom left corner represents areas with low similarity and low accuracy.

$$NS^i = \frac{\lambda_k^i - \lambda_{\min}^i}{\lambda_{\max}^i - \lambda_{\min}^i}, \tag{7}$$

$$NA_{aR}^i = \frac{aR_k^i - aR_{\min}^i}{aR_{\max}^i - aR_{\min}^i}, \tag{8}$$

$$NA_{aRMSE}^i = 1 - \frac{aRMSE_k^i - aRMSE_{\min}^i}{aRMSE_{\max}^i - aRMSE_{\min}^i}. \tag{9}$$

Figure 12 shows that for latent heat flux, the NLDAS-2 operational models span a broader range of the similarity-accuracy (S-A) space and the newer LSMs show a reduced coverage. For example, Noah28 encompasses a wider range of similarity with moderate accuracy levels. The similarity span is significantly reduced in Noah36 and NoahMP and they show an increase in the density of grid points with moderate accuracy and high similarity. The Mosaic LSM, on the other hand, shows considerable spread over areas with low/moderate accuracy. CLSM25 performance is similar to the Noah LSMs, with a large density of points in the

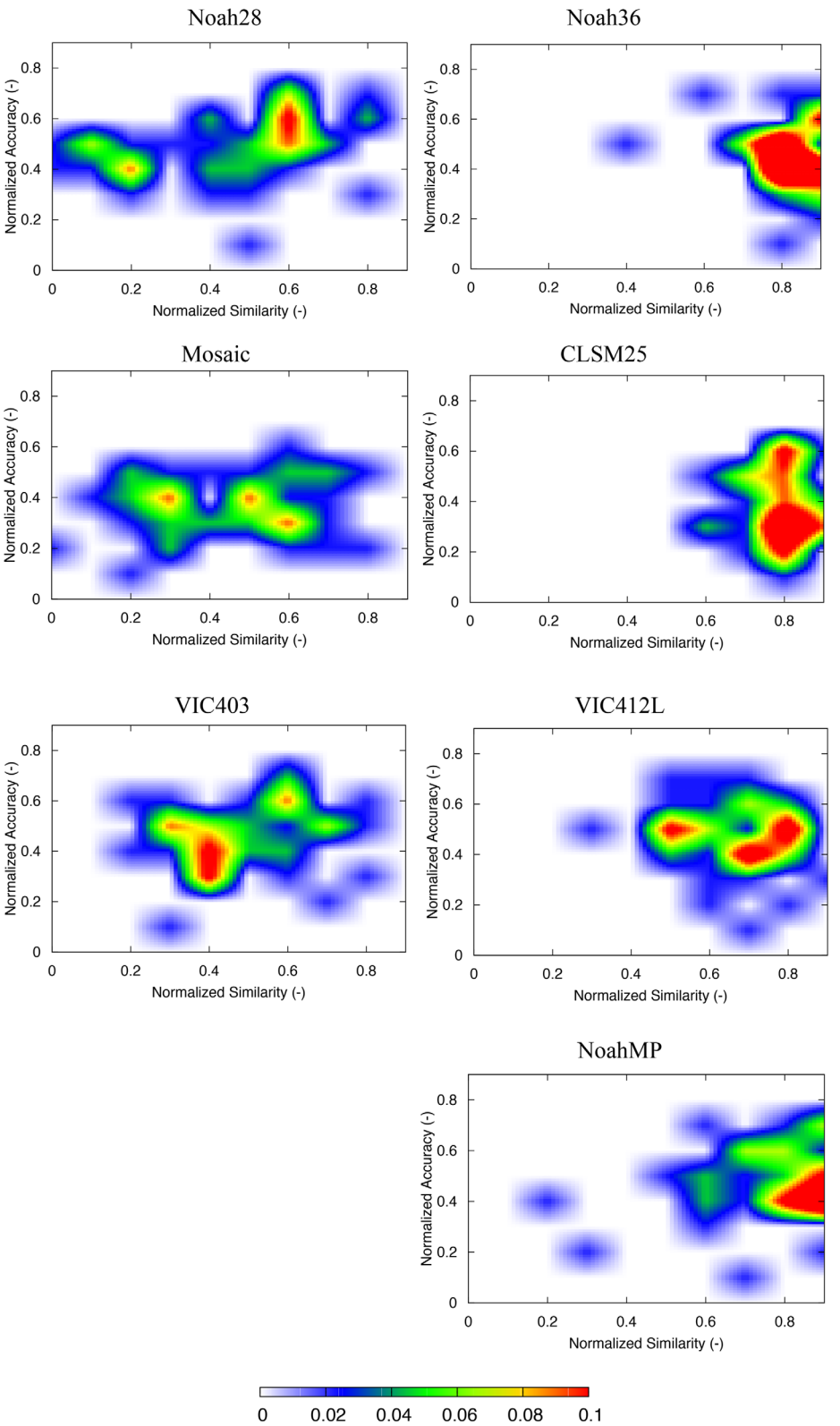


**Figure 12.** Density of grid points mapped as a function of the normalized similarity (*x* axis) and normalized accuracy (*y* axis) for latent heat flux.

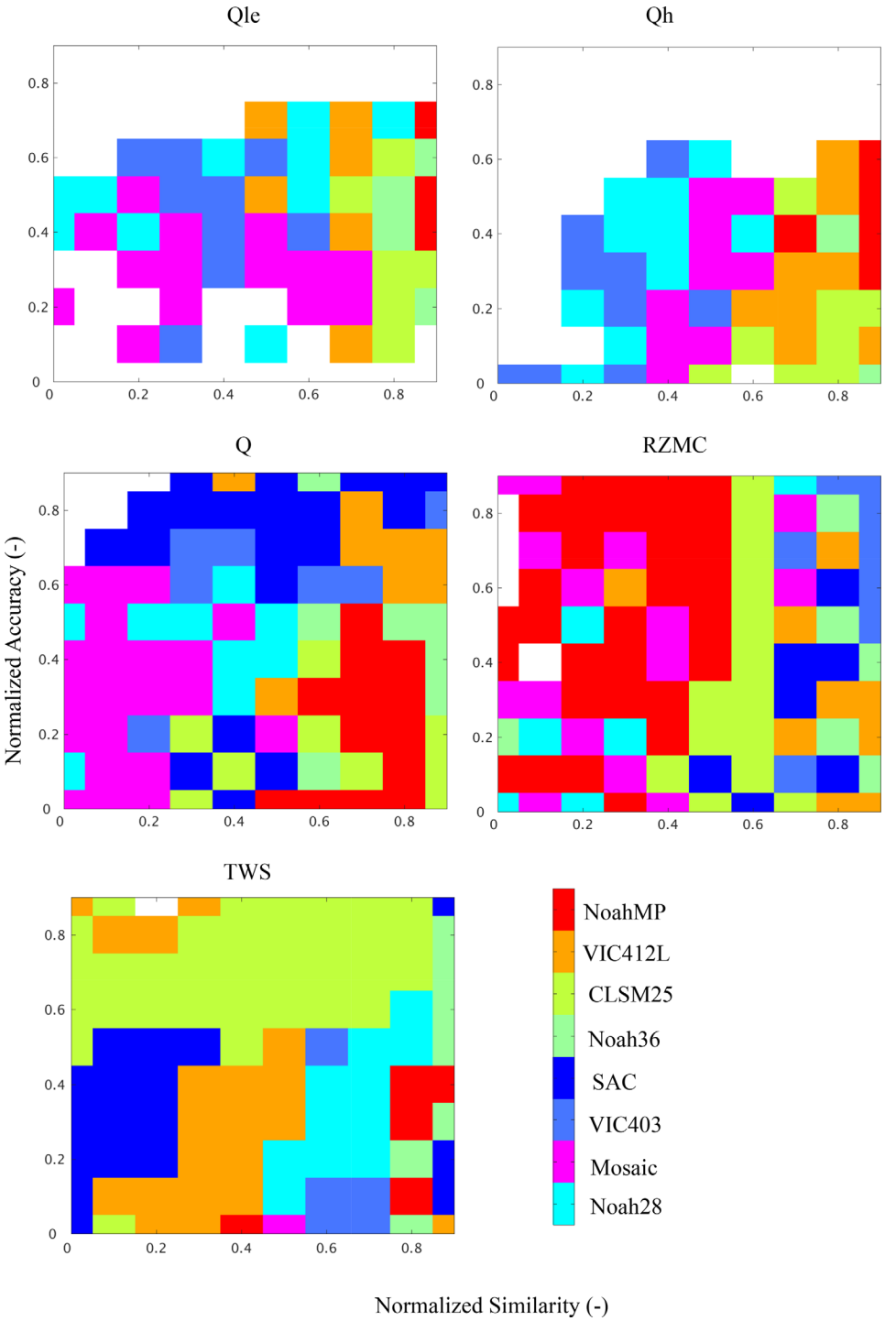


**Figure 13.** Model with the dominant span in each region of the normalized similarity (*x* axis) and normalized accuracy (*y* axis) for Qle, Qh, Q, RZMC, and TWS.

right part of the S-A space. The comparison of two different versions of the VIC model shows that the newer version of the model leads to an increased association with the common factor and an increase in the number of locations with improved accuracy.

Density maps similar to Figure 12 can be developed for Qh, Q, RZMC, and TWS (provided in the supporting information). The performance of the models in the S-A space is summarized in Figure 13, which shows the

model with the dominant span in each region of the S-A space for Qle, Qh, Q, RZMC, and TWS. In the Qle comparison, the NLDAS2 operational models (Noah28, Mosaic, and VIC) span the region with the low similarity and low to medium accuracy. As seen in Figure 12, Noah36, NoahMP, and VIC412L span the high similarity and high accuracy regions. Similar behavior is seen in the S-A space for Qh. In the RZMC comparison, Noah36 estimates have increased similarity to the common factor and increased accuracy relative to Noah28 (Noah28 mostly spans the low similarity and low accuracy region whereas Noah36 spans the region of high similarity). Compared to Mosaic, CLSM25 shows an increase in the density of points in the high accuracy space and an increased association with the common factor. There is significant reduction in the performance of VIC412L relative to that of VIC403. VIC412L mostly covers regions with low and medium accuracy and VIC403 encompasses regions with high accuracy though the similarity assessments of both models are comparable. The SAC model performance is similar to that of Noah36, spanning the regions of high similarity. Across all models, NoahMP shows the largest span in the similarity space with a large density of grid points with high accuracy. The S-A assessments for Q shows different trends relative to the Qle and RZMC comparisons. Generally, SAC shows the best performance, spanning the regions of high accuracy and low similarity. The VIC models also show high accuracy in their S-A space spans, whereas the Mosaic model encompasses the low similarity and the low accuracy regions. Among the models, CLSM25 shows the lowest skill and NoahMP and Noah36 span high similarity regimes. For TWS, CLSM25 shows the largest span in the similarity space with high accuracy, presumably because it is one of the LSMs that models groundwater storage. On the other hand, NoahMP (the other LSM with shallow groundwater formulations) spans the high similarity and low-medium accuracy space. The NLDAS operational models generally span the region of low to medium accuracy of the TWS S-A space. Note that in these plots, only the model with the highest density in a region is shown, though other LSMs may have comparable performance. Nevertheless, Figure 13 is helpful in indicating the relative strengths and weakness of the models in being able to produce skillful estimates without becoming too similar to the common factor.

## 6. Summary

Multiple land surface models are used in projects such as NLDAS to develop estimates of model prediction uncertainty and to increase the overall skill of the model predictions through the use of the ensemble. Due to differences in model physics and parameterizations, estimates of land surface processes from these models vary. As a result, the constituent models within an ensemble will have varying levels of similarity for different output variables. If the model outputs from different LSMs are too similar, their utility to the overall ensemble is low. Conversely, if the constituent models are deficient in their formulations and parameterizations, the individual model estimates could be very different from each other. In such instances, the outlier models may not be good choices for a multimodel ensemble. It is therefore important to understand the level of similarity and dissimilarity between models relative to the quality of their model predictions. This article provides such a quantitative evaluation of the similarity of model outputs from a suite of LSMs in the NLDAS-2 configuration.

A confirmatory factor analysis that describes the variability across the constituent models within the ensemble through the development of unobserved latent, common factors is used for assessing model similarity across the models. In addition to the four operational NLDAS-2 models, four additional models that represent newer versions and modern advancements in the model physics are considered in this analysis. All models are forced with the same meteorology and evaluated over an 11 year time period from 2002 to 2012. The analysis in the article focuses on five variables (latent heat flux, sensible heat flux, total runoff, root zone soil moisture, and snow water equivalent) that represent some of the key components of the terrestrial water and energy budgets. For the sake of uniformity, all analyses are conducted in the anomaly space of the variable, ignoring the skill of the mean seasonal cycles. The main output of the factor analysis is the factor loading for each model, which is an estimate of how well the estimates from that model agrees with the factor common to all the LSMs. Factor loading values close to zero indicate that the model estimate has a weak association with the common factor, whereas departures from zero indicate that the model estimate is strongly associated with the common factor.

An assessment of the intermodel similarity for each variable is first developed by comparing the average anomaly R among the 8 LSMs at the daily time scale. The Q estimates were most dissimilar whereas RZMC,

SWE, and TWS estimates showed high degree of similarity. A stratification of the inter model similarity based on Köppen-Geiger climate classification zones indicated varying levels of similarity and dissimilarity for different variables. Generally, the models showed greater agreement among themselves when the contributing physical mechanisms did not involve multiple processes. For example, the runoff estimates showed greater agreement in arid zones, where soil moisture is the primary controlling factor in runoff generation. Similarly, the surface fluxes showed greater level of agreement in the warm temperate zones, and SWE estimates were more similar in the cold zones. Temporal stratification of the intermodel similarity also indicated similar trends. The surface fluxes showed greater agreement during the warm time periods whereas SWE estimates were more similar across the models during the cold season.

The factor loadings help to characterize modes of similarity among ensemble members. Generally, in the comparison of the factor loadings, the newer models show stronger association with the common factor whereas larger spatial variability in the factor loadings is seen with the NLDAS-2 operational models. For surface heat fluxes and total runoff, the new LSMs and the operational LSMs fall into two distinct groups with the former showing strong association with the common factor and the latter showing weak association. For RZMC, SWE, and TWS, no such obvious contrasts are seen across the models. A map of the contribution of the constituent models toward the ensemble based on their level of association with the common factor shows large variability with different LSMs demonstrating dominant contributions based on the variable of interest. The analysis of the factor loadings at the monthly time scale shows that the level of similarity among the models significantly increases at longer time scales.

A quantitative assessment of model similarity alone is not sufficient to determine a model's usefulness to the ensemble. The article also presents a trade-off analysis with a simultaneous assessment of model accuracy and similarity. Overall, the NLDAS-2 operational models display larger spread in the similarity-accuracy space. The new LSMs encompass a narrower region of the S-A space with increased similarity and accuracy, with some exceptions. For example, the RZMC estimates are less accurate with higher levels of similarity with the common factor for the new version of the VIC model.

Overall, the analysis presented in this article is a useful way to assess the contribution of the constituent models within an ensemble. This is particularly relevant for multimodel projects such as NLDAS where the ensemble is used for a variety of applications at different time scales. As new models and model configurations are introduced, a simultaneous assessment of similarity and accuracy can be conducted to benchmark their utility. Acceptable benchmark thresholds in the similarity and accuracy space can be established as formal criteria for the selection of a model within an ensemble. Such benchmarks can also be used to guide future model development, which has traditionally been driven by accuracy requirements alone. For example, it is conceivable that instead of employing multiple LSMs, sufficient spread in the similarity with high accuracy can be achieved by sampling the model parameter space of a single model. This study serves as a benchmark for such future research.

It must be clarified that the article focuses on an information-based similarity analysis. A similar distance-based similarity analysis can be developed by focusing on a direct comparison of the variables from the LSMs (rather than a comparison of anomalies) and by using distance-based metrics such as mean squared error rather than correlation. In addition, the composition of the ensemble also affects the results. For example, if only the NLDAS operational models are used in the similarity quantification, the results may indicate that they show strong association with the common factor across those four LSMs. The results also show that factors such as climate regime, variable and time scale of interest, model similarity and accuracy must be considered in the utility assessment and the development of multimodel ensembles.

**Acknowledgments**
Funding for this work was provided by the NOAA's Climate Program Office (MAPP program). Computing was supported by the resources at the NASA Center for Climate Simulation. The NLDAS-2 forcing data used in this effort were acquired as part of the activities of NASA's Science Mission Directorate and are archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC). We also gratefully acknowledge the use of the AmeriFlux data provided on the website ameriflux.lbl.gov.

## References

Anderson, E. (1973). *National Weather Service River Forecast System-Snow Accumulation and Ablation Model* (*NOAA Tech. Memo. NWS Hydro-17*). Washington, DC: US National Weather Service.

Ball, J. T., Woodrow, I. E., & Berry, J. A. (1987). *A model predicting stomatal conductance and its contribution to the control of photosynthesis under different environmental conditions* (pp. 221–224). Dordrecht, the Netherlands: Springer. https://doi.org/10.1007/978-94-017-0519-6_48

Barlage, M., Chen, F., Tewari, M., Ikeda, K., Gochis, D., Dudhia, J., . . . Mitchell, K. (2010). Noah land surface model modifications to improve snowpack prediction in the Colorado Rocky Mountains. *Journal of Geophysical Research*, *115*, D22101. https://doi.org/10.1029/2009JD013470

Burnash, R., Ferral, R., & McGuire, R. (1973). *A generalized streamflow simulation system: Conceptual models for digital computer* (Tech. Rep.). Sacramento, CA: Joint Federal-State River Forecast Center.

Cai, L. (2012). Latent variable modeling. *Shanghai Arch Psychiatry*, *24*(2), 118–120. https://doi.org/10.3969/j.issn.1002-0829.2012.02.010
Chen, F., & Dudhia, J. (2001). Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Monthly Weather Review*, *129*(4), 569–585.
Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H.-L., Koren, V., . . . Betts, A. (1996a). Modeling of land surface evaporation by four schemes and comparison with fife observations. *Journal of Geophysical Research*, *101*(D3), 7251–7268. https://doi.org/10.1029/95JD02165
Chen, F., Mitchell, K., Schaake, J., Xue, Y. K., Pan, H., Koren, L. V., . . . Betts, A. (1996b). Modeling of land surface evaporation by four schemes and comparison with fife observations. *Journal of Geophysical Research*, *101*(D3), 7251–7268.
Christensen, W., & Sain, S. (2012). Latent variable modeling for integrating output from multiple climate models. *Mathematical Geosciences*, *44*, 395–410. https://doi.org/10.1007/s11004-011-9321-1
Dirmeyer, P., Gao, X., Zhao, M., Guo, Z., Oki, T., & Hanasaki, N. (2006). GSWP-2: Multimodel analysis and implications for our perception of the land surface. *Bulletin of the American Meteorological Society*, *87*, 1381–1397. https://doi.org/10.1175/BAMS-87-10-1381
Ducharne, A., Koster, R., Suarez, M., Stieglitz, M., & Kumar, P. (2000). A catchment-based approach to modeling land surface processes in a general circulation model: 2. Parameter estimation and model demonstration. *Journal of Geophysical Research*, *105*(D20), 24823–24838.
Ek, M., Mitchell, K., Yin, L., Rogers, P., Grunmann, P., Koren, V., . . . Tarpley, J. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research*, *108*(D22), 8851. https://doi.org/10.1029/2002JD003296
Gao, H., Tang, Q., Shi, X., Zhu, C., Bohn, T. J., Su, F., . . . Wood, E. F. (2010). Water budget record from Variable Infiltration Capacity (VIC) model. In *Algorithm theoretical basis document* (Tech. Rep.). Seattle, WA: Department of Civil Engineering, University of Washington.
Giorgi, F., & Mearns, L. (2002). Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method. *Journal of Climate*, *15*, 1141–1158.
Hansen, M., DeFries, R., Townshend, J., & Sohlberg, R. (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing*, *21*(6), 1331–1364.
Hattie, J., & Fraser, C. (1988). The constraining of parameters in a restricted factor analysis. *Applied Psychological Measurement*, *12*, 155–162.
Jarvis, P. G. (1976). The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *273*(927), 593–610. https://doi.org/10.1098/rstb.1976.0035
Koppen, W. (2011). The thermal zones of the earth according to the duration of hot, moderate and cold periods and to the impact of heat on the organic world. *Meteorologische Zeitschrift*, *20*, 351–360. https://doi.org/10.1127/0941-2948/2011/105
Koster, R., & Suarez, M. (1996). *Energy and water balance calculations in the Mosaic LSM* (Tech. Rep. 104606). Washington, DC: NASA.
Koster, R. D., & Suarez, M. J. (1992). Modeling the land surface boundary in climate models as a composite of independent vegetation stands. *Journal of Geophysical Research*, *97*(D3), 2697–2715.
Koster, R. D., Suarez, M. J., Ducharne, A., Stieglitz, M., & Kumar, P. (2000). A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. *Journal of Geophysical Research*, *105*(D20), 24809–24822.
Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3), 259–263. https://doi.org/10.1127/0941-2948/2006/0130
Kumar, S., Peters-Lidard, C. D., Mocko, D., Reichle, R., Liu, Y., Arsenault, K. R., . . . Cosh, M. (2014). Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. *Journal of Hydrometeorology*, *15*, 2446–2469. https://doi.org/10.1175/JHM-D-13-0132.1
Kumar, S., Reichle, R., Harrison, K., Peters-Lidard, C., Yatheendradas, Y., & Santanello, J. (2012). A comparison of methods for a priori bias correction in soil moisture data assimilation. *Water Resources Research*, *48*, W03515. https://doi.org/10.1029/2010WR010261
Kumar, S., Zaitchik, B. F., Peters-Lidard, C. D., Rodell, M., Reichle, R., Li, B., . . . Ek, M. (2016). Assimilation of gridded GRACE terrestrial water storage estimates in the North American Land Data Assimilation System. *Journal of Hydrometeorology*, *17*(7), 1951–1972.
Kumar, S. V., Peters-Lidard, C. D., Tian, Y., Houser, P. R., Geiger, J., Olden, S., . . . Sheffield, J. (2006). Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modeling and Software*, *21*, 1402–1415.
Landerer, F., & Swenson, S. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resources Research*, *48*, W04531. https://doi.org/10.1029/2011WR011453
Li, M., Chen, X., Li, X., Ma, B., & Vitanyi, P. (2004). The similarity metric. *IEEE Transactions on Information Theory*, *50*(12), 3250–3264.
Liang, X., Lettenmaier, D., Wood, E., & Burges, S. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, *99*(D7), 14415–14428. https://doi.org/10.1029/94JD00483
Lin, D. (1998). An information-theoretic definition of similarity. In *ICML '98 proceedings of the fifteenth international conference on machine learning* (pp. 296–304). San Francisco, CA: Morgan Kaufmann Publishers Inc.
Miller, D., & White, R. (1998). A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. *Earth Interactions*, *2*, 1–26.
Mitchell, K., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., . . . Bailey, A. A. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, *109*, D07S90. https://doi.org/10.1029/2003JD003823
Mo, K. C., Chen, L.-C., Shukla, S., Bohn, T. J., & Lettenmaier, D. P. (2012). Uncertainties in North American Land Data Assimilation Systems over the contiguous United States. *Journal of Hydrometeorology*, *13*(3), 996–1009. https://doi.org/10.1175/JHM-D-11-0132.1
Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M., & Stainforth, D. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*, 768–772.
Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., . . . Xia, Y. (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*, *116*, D12109, https://doi.org/10.1029/2010JD015139
Palmer, T., Doblas-Reyes, F., Hagedorn, R., & Weisheimer, A. (2005). Probabilistic prediction of climate using multi-model ensembles: From basics to applications. *Philosophical Transactions of the Royal Society B*, *360*, 1991–1998.
Pan, M., Sheffield, J., Wood, E. F., Mitchell, K. E., Houser, P. R., Schaake, J. C., . . . Tarpley, J. D. (2003). Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model simulated snow water equivalent. *Journal of Geophysical Research*, *108*(D22), 8850. https://doi.org/10.1029/2003JD003994
Peters-Lidard, C. D., Houser, P. R., Tian, Y., Kumar, S. V., Geiger, J., Olden, S., . . . Sheffield, J. (2007). High-performance Earth system modeling with NASA/GSFC's Land Information System. *Innovations in Systems and Software Engineering*, *3*(3), 157–165.
Pitman, A., & Henderson-Sellers, A. (1998). Recent progress and results from the project for the intercomparison of land surface parameterization schemes. *Journal of Hydrology*, *212–213*, 128–135.
Rastetter, E. B. (1996). Validating models of ecosystem response to global change. *BioScience*, *46*(3), 190–198.

Reichle, R., Koster, R., Lannoy, G. D., Forman, B., Liu, Q., & Mahanama, S. (2011). Assessment and enhancement of MERRA land surface hydrology estimates. *Journal of Climate*, *24*, 6322–6338. https://doi.org/10.1175/JCLI-D-10-05033.1

Richards, L. A. (1931). Capillary conduction of liquids through porous mediums. *Physics*, *1*(5), 318–333.

Robock, A., Luo, L., Wood, E. F., Wen, F., Mitchell, K. E., Houser, P. R., . . . Crawford, K. C. (2003). Evaluation of the North American Land Data Assimilation System over the southern Great Plains during the warm season. *Journal of Geophysical Research*, *108*(D22), 8846. https://doi.org/10.1029/2002JD003245

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., . . . Toll, D. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, *85*(3), 381–394.

Schaefer, G., Cosh, M., & Jackson, T. (2007). The USDA Natural Resources Conservation Service soil climate analysis network (SCAN). *Journal of Atmospheric and Oceanic Technology*, *24*, 2073–2077.

Schwalm, C. R., Huntzinger, D. N., Fisher, J. B., Michalak, A. M., Bowman, K., Ciais, P., . . . Zeng, N. (2015). Toward "optimal" integration of terrestrial biosphere models. *Geophysical Research Letters*, *42*, 4418–4428. https://doi.org/10.1002/2015GL064002

Sellers, P. J., Mintz, Y., Sud, Y. C., & Dalcher, A. (1986). A Simple Biosphere Model (SIB) for use within general circulation models. *Journal of the Atmospheric Sciences*, *43*(6), 505–531. https://doi.org/10.1175/1520-0469(1986)043<0505:ASBMFU>2.0.CO;2

Sheffield, J., Pan, M., Wood, E. F., Mitchell, K. E., Houser, P. R., Schaake, J. C., . . . Ramsay, B. H. (2003). Snow process modeling in the North American Land Data Assimilation System (NLDAS): 1. Evaluation of model-simulated snow cover extent. *Journal of Geophysical Research*, *108*(D22), 8849. https://doi.org/10.1029/2002JD003274

Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *24*(4), 35–43.

Tapley, B., Bettadpur, S., Watkins, M., & Reigber, C. (2004). The Gravity Recovery and Climate Experiment: Mission overview and early results. *Geophysical Research Letters*, *31*, L09607. https://doi.org/10.1029/2004GL019920

Tebaldi, C., Smith, R., Nychka, D., & Mearns, L. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multi-model ensembles. *Journal of Climate*, *18*, 1524–1540.

Tryon, R. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Ann Arbor, MI: Edwards Brothers.

Wang, Z., Zeng, X., & Decker, M. (2010). Improving snow processes in the Noah land model. *Journal of Geophysical Research*, *115*, D20108, https://doi.org/10.1029/2009JD013761

Wei, H., Xia, Y., Mitchell, K. E., & Ek, M. B. (2013). Improvement of the Noah land surface model for warm season processes: evaluation of water and energy flux simulation. *Hydrological Processes*, *27*(2), 297–303. https://doi.org/10.1002/hyp.9214

Xia, Y., Cosgrove, B., Mitchell, K., Peters-Lidard, C., Ek, M., Kumar, S., . . . Wei, H. (2016a). Basin-scale assessment of the land surface energy budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *Journal of Geophysical Research: Atmospheres*, *121*, 196–220. https://doi.org/10.1002/2015JD023889

Xia, Y., Cosgrove, B. A., Ek, M. B., Sheffield, J., Luo, L., Wood, E. F., . . . Xia, Y. (2013). Overview of the North American Land Data Assimilation System (NLDAS). In *Land surface observation, modeling and data assimilation* (pp. 337–377). Singapore, Singapore: World Scientific. https://doi.org/10.1142/9789814472616_0011

Xia, Y., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Brewer, M., . . . Luo, L. (2016b). Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *Journal of Geophysical Research: Atmospheres*, *121*, 2750–2779. https://doi.org/10.1002/2015JD023733

Xia, Y., Ek, M., Wei, H., & Meng, J. (2012a). Comparative analysis of relationships between NLDAS-2 forcings and model outputs. *Hydrological Processes*, *26*(3), 467–474. https://doi.org/10.1002/hyp.8240

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., . . . Lohmann, D. (2012c). Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *Journal of Geophysical Research*, *117*, D03110, https://doi.org/10.1029/2011JD016051

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., . . . Mocko, D. (2012b). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research: Atmospheres*, *117*, D03110, https://doi.org/10.1029/2011JD016048

Xia, Y., Mocko, D., Huang, M., Li, B., Rodell, M., Mitchell, K. E., . . . Ek, M. B. (2017). Comparison and assessment of three advanced land surface models in simulating terrestrial water storage components over the United States. *Journal of Hydrometeorology*, *18*, 625–649. https://doi.org/10.1175/JHM-D-16-0112.1

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., . . . Xia, Y. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research*, *116*, D12110. https://doi.org/10.1029/2010JD015140