JAMES | Journal of Advances in Modeling Earth Systems

AGU ADVANCING EARTH AND SPACE SCIENCE

RESEARCH ARTICLE
10.1029/2020MS002132

Key Points:

- A novel framework of falsification-oriented signature-based evaluation is proposed
- The framework is demonstrated to reveal the regions of inconsistencies between model predictions and observations
- The identified regions coincide with the sites where both theoretical modeling advancements and new observational data (e.g., from the Critical Zone Observatories) have emerged

Supporting Information:

- Supporting Information S1
- Data Set S1

Correspondence to:
Z.-L. Yang,
liang@jsg.utexas.edu

Citation:
Zheng, H., Yang, Z.-L., Lin, P., Wu, W.-Y., Li, L., Xu, Z., et al. (2020). Falsification-oriented signature-based evaluation for guiding the development of land surface models and the enhancement of observations. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002132. https://doi.org/10.1029/2020MS002132

Received 19 APR 2020
Accepted 8 NOV 2020
Accepted article online 12 NOV 2020

Falsification-Oriented Signature-Based Evaluation for Guiding the Development of Land Surface Models and the Enhancement of Observations

Hui Zheng[1], Zong-Liang Yang[2], Peirong Lin[2,3], Wen-Ying Wu[2], Lingcheng Li[2], Zhongfeng Xu[1], Jiangfeng Wei[4], Long Zhao[5], Qingyun Bian[1], and Shu Wang[6]

[1]Key Laboratory of Regional Climate-Environment Research for Temperate East Asia, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China, [2]Department of Geological Sciences, The John A. and Katherine G. Jackson School of Geosciences, University of Texas at Austin, Austin, TX, USA, [3]Now at Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA, [4]Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters/Key Laboratory of Meteorological Disaster, Ministry of Education/International Joint Research Laboratory on Climate and Environment Change, Nanjing University of Information Science and Technology, Nanjing, China, [5]School of Geographical Sciences, Southwest University, Chongqing, China, [6]State Key Laboratory of Operation and Control of Renewable Energy and Storage Systems, China Electric Power Research Institute, Beijing, China

Abstract We develop a novel framework for rigorously evaluating land surface models (LSMs) against observations by recognizing the asymmetry between verification- and falsification-oriented approaches. The former approach cannot completely verify LSMs even though it exhausts every case of consistency between the model predictions and observations, whereas the latter only requires a single case of inconsistency to reveal that there must be something wrong. We argue that it is such an inconsistency that stimulates further development of the models and enhancement of the observations. We therefore propose a falsification-oriented signature-based evaluation framework to identify cases of inconsistency between model predictions and observations by extracting signatures based on a set of key assumptions. We apply this framework to evaluate an ensemble of simulations from the Noah-MP LSM against observations over the continental United States under the three assumptions of water mass conservation, no lateral water flow, and a sufficiently long period of time. Regions showing inconsistencies between the Noah-MP ensemble simulations and the observations are located in the western mountainous areas, the Yellowstone river basin, the lower Floridan aquifer, the Niobrara river basin at the north tip of the Ogallala aquifer, and the basins downstream of the Balcones fault zones in Texas. These regions coincide with the sites where both advances in theoretical modeling and new observational data (e.g., from the Critical Zone Observatories) have emerged.

Plain Language Summary We propose a framework for locating regions that require substantial efforts in modeling and observations. The framework is based on the fact that a single case of inconsistency between model predictions and observations always indicates that there must be something wrong. Such an consistency can stimulate and guide further development of models and enhancement of observations. An application over the continental United States shows that the identified regions encompass the areas where both theoretical modeling advancements and observational data from the Critical Zone Observatories have emerged. The results also suggest the need to extend the observatories over the Balcones fault zones in Texas and the Floridan aquifer in Florida.

1. Introduction

1.1. Difficulties in Verification

Rigorous evaluations of land surface models (LSMs) are necessary for guiding the developments and applications of the models, but the widely adopted verification-oriented evaluations do not fully meet the intended objectives. There are always known knowns, known unknowns, and unknown unknowns. Models represent the known knowns well but suffer from known unknowns (e.g., the uncertain parameterization of known

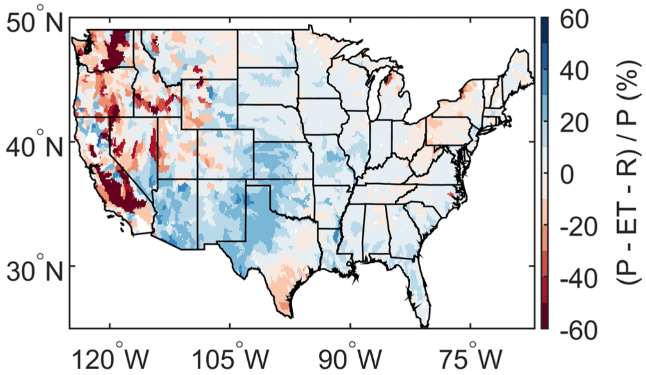

Figure 1. Water budget imbalance among the NLDAS precipitation (*P*), FLUXNET MTE evapotranspiration (*ET*), and USGS runoff (*R*) at each HUC8 basin over the CONUS. The water budget imbalance is calculated as the difference between the 30-year (1982–2011) average annual precipitation and the sum of the annual evapotranspiration and runoff. Detailed descriptions of the data sets are given in section 4.1.

land surface processes and known subscale heterogeneity) and unknown unknowns (e.g., unrepresented boundary conditions, unresolved subscale processes, and unknown biogeochemical processes). Solving the model equations requires that the modeled system has to be closed, but the existence of unknown unknowns constantly challenges this assumption. As a consequence, rigorous verification of LSMs is impossible (Oreskes et al., 1994): Even if a model prediction is consistent with all the known observations in all the known criteria, the model still cannot be completely verified. The interpretations of the verification-oriented evaluation results are inevitably inconclusive (Hill et al., 2017; Kirchner, 2006).

Observations also involve known unknowns and unknown unknowns; the errors in existing large-scale land surface water budget observations can be shown to be significant. Assuming a sufficiently long time and no lateral flow of groundwater (detailed in section 3.1), the long-term average precipitation should be reasonably balanced by evapotranspiration plus runoff (Kauffeldt et al., 2013; Thornthwaite, 1948; Wilm et al., 1944). However, the analysis of available 30-year multisource data sets over the continental United States (CONUS) shows that the magnitude of the imbalance is troubling (Figure 1): >10% of precipitation in most of the CONUS and >60% of precipitation in the West. Such an imbalance can be attributed to unmeasured lateral flow of groundwater and the observational errors in precipitation (Adam et al., 2006; Henn et al., 2018), evapotranspiration (Long et al., 2014), and runoff (Wilby et al., 2017). The observational errors stem from insufficient spatiotemporal sampling, instrumental limitations, site changes, human activities, and erroneous data archiving and postprocessing, thereby making it infeasible to close the terrestrial water budget (Pan et al., 2012; Sahoo et al., 2011; Sheffield et al., 2009) across various databases (Kauffeldt et al., 2013).

Observational errors greatly complicate evaluations. If multiple data sets are not physically consistent (e.g., Figure 1), the observations of different water budget components may present conflicting information (Beven & Westerberg, 2011; Kauffeldt et al., 2013; Pan et al., 2012; Sheffield et al., 2009). The evaluation of a model using such kind of data sets would be unavoidably controversial. In general, observations are used to drive models and to evaluate model predictions. If the driving observations contain errors, then even a perfect model can generate problematic predictions. If the observations used for the evaluations contain errors, then the consistency between the observations and the model predictions should be a sign of modeling errors. In either case, the consistency between model predictions and observations should not be considered as a solid indicator of accurate predictions.

In the recognization of the observational and model prediction errors, the likelihood of a model being true is often estimated. However, the validation of this approach is conditional on a subjective assumption: Unknown unknowns can be neglected. It may be reasonable to neglect unknown unknowns at a well-controlled reference site, but in the strictest sense, unknown unknowns exist at all times. When unknown unknowns cannot be neglected, then the estimation of likelihood inevitably involves an infinite logical regression (Popper, 2002). A priori knowledge about the truth is always necessary (e.g., an a priori value, an a priori error between the initial guess and the truth, or an a priori error distribution), but our prior knowledge always involves unknown unknowns, which are logically impossible to know in a priori.

1.2. Falsification Overcomes the Difficulties of Verification

As mentioned above, rigorous verification-oriented evaluation is impossible as a result of the existence of unknown unknowns. The difficulty is deeply rooted in our epistemological foundations, impeding our advances in scientific understanding and modeling capability. How, then, can a rigorous evaluation be performed? Signature-based (Gupta et al., 2008) hypothesis testing (Beven, 2001, 2018) subject to falsification (Popper, 2002) appears to be a viable approach.

In an attempt to apply this approach, section 2 introduces several key aspects of falsification-oriented signature-based evaluation, showing how the difficulties in verification can be avoided. Section 3 proposes a practical framework for evaluating long-term land surface water budgets. As an application of the framework, section 4 presents an experiment over the CONUS. Section 5 gives our results, which are discussed in section 6, followed by our conclusions in section 7.

2. The Falsification-Oriented Signature-Based Evaluation Paradigm

Verification is "an assertion or establishment of truth" (Refsgaard & Henriksen, 2004), whereas falsification is an assertion or establishment of false. Verification- and *falsification*-oriented evaluations aim to test the consistency and *inconsistency*, respectively, between model predictions and observations.

As discussed in section 1, the impossibility of rigorous verification has long been recognized (Oreskes et al., 1994; Popper, 2002). The inclusiveness of verification-oriented evaluations plays a notable part in the coexistence of numerous, but different, models. The falsification-oriented approach has therefore attracted increasing attention (Baker, 2017; Beven, 2018; Blöschl, 2017; Linde, 2014; McKnight, 2017; Neuweiler & Helmig, 2017; Pfister & Kirchner, 2017). However, these opinion papers have overlooked the asymmetry between verification and falsification, and there is a lack of practical frameworks and successful applications of this approach.

2.1. Asymmetry Between Falsification and Verification

It is essential to recognize the asymmetry between falsification and verification: Every consistency between model predictions and observations cannot completely verify the model as a result of the existence of unknown unknowns, whereas one single inconsistency always indicates that there must be something wrong. The inconsistency is a solid indicator of errors of any kind, regardless of whether they stem from known unknowns or unknown unknowns.

Popper (2002) summarized the philosophy of falsification as "all our knowledge grows only by correcting our mistakes." "Learning from mistakes" (Beven, 2001, 2018) is a defining characteristic of falsification. With the falsification-oriented approach, the reasons for the inconsistencies (mistakes) between the model predictions and the observations are investigated. Better predictions and observations are the result of addressing these "mistakes" (Niu et al., 2005, 2007, 2011; Yang et al., 2011). Son and Sivapalan (2007) provided an excellent example of how models can be improved by learning from "wrong" predictions. However, with the verification-oriented approach, "good" predictions are selected by showing their consistencies with the observations. The likelihood of a scientific hypothesis being true is often estimated by using the "good" predictions that are consistent with the observations. The "wrong" predictions, which are rich in information about how to improve the models and observations further, are largely abandoned.

2.2. The "Try-Fail-Refine" Strategy for Exploring Unknowns

The inconsistency between model predictions and observations may stem from known unknowns and unknown unknowns. Falsification-oriented evaluation can detect unknown unknowns by considering known unknowns. Figure 2 shows that a comparison between model predictions and observations results in three possible outcomes: (1) not falsified, (2) partially falsified, and (3) fully falsified. In the case of the fully falsified situation, if the known unknowns have been carefully considered, then unknown unknowns must exist.

The consideration of known unknowns (Beven, 1993, 2004, 2009; Beven & Binley, 1992, 2014; Binley et al., 1991; Sivapalan et al., 2003) can be iteratively refined with a "try-fail-refine" strategy. In an experiment, known unknowns are specified when the observations, models (i.e., parameterizations and parameters), and evaluation methods are chosen. If the model predictions are asserted as inconsistent with the observations, then it is possible that too few known unknowns have been considered. The experiment can be refined by considering more known unknowns. The refinement can be performed in sequence, as demonstrated by Son and Sivapalan (2007). Each iteration provides additional insights into the unknowns. If all the feasible known unknowns have been considered, then there must be unknown unknowns. A fully falsified test of this case is a strong stimulation of pursuing new theories, modeling new processes, and observing new phenomena (Popper, 2002; Sherwood, 2011).

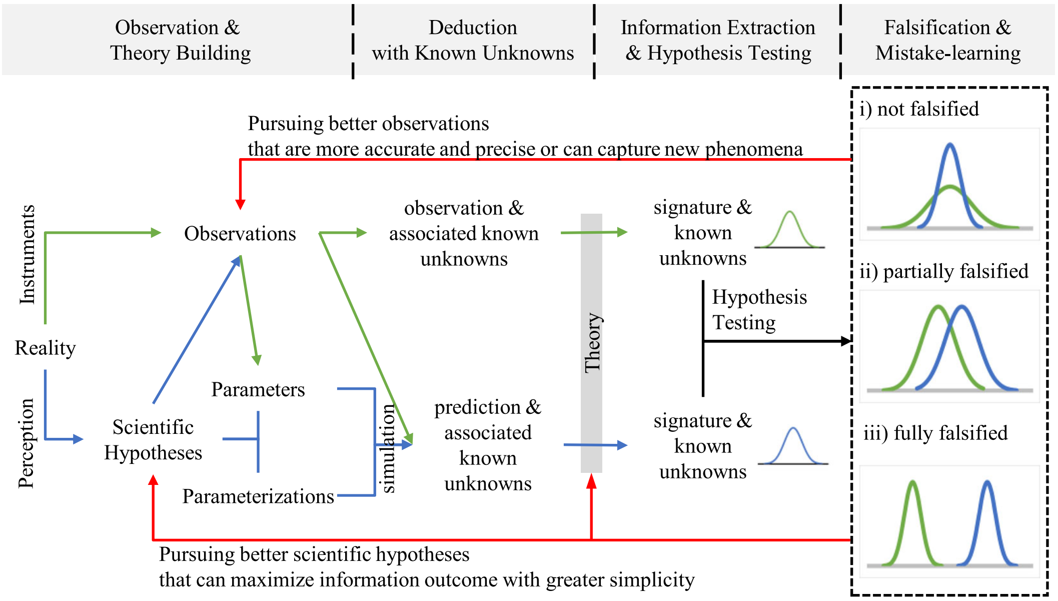

Figure 2. Diagram showing the falsification-oriented signature-based evaluation paradigm. The paradigm is divided into several phases, as shown at the top. The hypothesis is whether the prediction signature is inconsistent with the observation signature. The three outcomes of the falsification-oriented hypothesis testing are shown in the dashed box. Further details can be found in section 2.

2.3. Scientific Hypotheses Are Not Approximations of the Truth

As discussed in section 1, the existence of unknown unknowns means that any a priori estimates or approximations of the truth cannot be reliable. They should be avoided entirely in rigorous evaluations.

The falsification-oriented evaluation is built on the test of scientific hypotheses. The term “scientific hypotheses” is used here to exclude mathematical and logical hypotheses. It is worth noting that scientific hypotheses are not approximations of the truth. Instead, they represent our growing understanding of the truth and compete to produce fewer “mistakes” (i.e., fewer inconsistencies between the predictions and the observations). Figure 2 shows that scientific hypotheses are involved in models, observations, and evaluation methods.

A numerical LSM is a scientific hypothesis. A model assembles a set of parameterizations and parameters. A parameterization is a scientific hypothesis about a relationship between different natural phenomena (e.g., precipitation, soil moisture, and runoff). A parameter is an adjustable coefficient in the relationship (Clark et al., 2011) and represents a scientific hypothesis about the properties of natural systems. In practice, parameters can be assigned hypothetically, either with observations or with other scientific hypotheses such as pedotransfer functions (Van Looy et al., 2017). Model predictions are made by numerically solving the mathematical equations that represent parameterizations (e.g., infiltration and runoff), parameters (e.g., soil hydraulic conductivity), and observations (e.g., precipitation and solar radiation). The solving process is a form of rigorous deduction.

“Observation always involves theory” (i.e., a set of scientific hypotheses) (Hubble, 2013). Observations of terrestrial water fluxes inevitably involve scientific hypotheses about the relationships between the phenomena intended to be observed (e.g., streamflow or evapotranspiration) and the phenomena that can be observed (e.g., water level or soil moisture). Scientific hypotheses also have to be made about subscale processes and heterogeneity, which always contain known unknowns and unknown unknowns. If a model prediction is inconsistent with the observations, then errors may stem from the scientific hypotheses that underlie the observations.

2.4. Signature-Based Evaluations: Evaluating the Evaluation Methods

Evaluation methods also involve scientific hypotheses. Gupta et al. (2008) proposed signature-based evaluations, in which the signatures of model predictions and observations are compared. A signature is information extracted from the model predictions and observations based on some theories (i.e., a set of scientific hypotheses) and is used to measure functional behavior (Wagener & Montanari, 2011). The scientific hypotheses of evaluation methods or signature extraction should also be tested.

We propose that a signature should be rigorously deduced from model predictions and observations based on a set of scientific hypotheses. With a rigorous deduction, the logical rule that a false conclusion must have false premises will be useful. If model predictions and observations disagree in their signatures, the inconsistency can stem from (1) the errors in the observations, (2) the model that gives the predictions (i.e., parameterizations and parameters), and (3) the invalidation of the evaluation method (i.e., signature extraction).

3. Framework for Evaluating Long-Term Land Surface Water Budgets

A practical framework for evaluating long-term land surface water budgets is proposed. The framework extracts signatures based on three scientific hypotheses: (1) The mass conservation of water at the land surface is satisfied; (2) there is no horizontal water exchange between adjacent basins/grids; and (3) the length of time is sufficiently long.

Three facets of the framework are described. First, we introduce the three scientific hypotheses of the framework and illustrate the framework using a diagram. Second, we explore how the signatures are extracted based on the three scientific hypotheses. Third, we describe the rules of falsifying the consistency between model predictions and observations.

3.1. Three Scientific Hypotheses and a Diagram

For a control volume at the land surface (e.g., catchments or subcatchment units), water inputs, outputs, and storage change should be subject to the physical principle of mass conservation (the first scientific hypothesis) (Eagleson, 1978; Reggiani et al., 2000, 2001; Reggiani & Schellekens, 2003). With the assumption that there is no horizontal water exchange between adjacent control volumes (the second scientific hypothesis), a lumped water balance equation (Beven et al., 2011) can be classically written as follows:

$$p = et + r + ds \tag{1}$$

where p is the precipitation rate (kg m^{-2} s^{-1}), et is the evapotranspiration rate (kg m^{-2} s^{-1}), r is the runoff rate (kg m^{-2} s^{-1}), and ds is the internal change rate of the storage at all points and at all levels in the control volume (kg m^{-2} s^{-1}).

Equation 1 can be integrated over a period of time Δt (s):

$$\begin{aligned} &P = ET + R + \frac{\Delta S}{\Delta t} \\ &P\Delta t = \int_{t=0}^{\Delta t} p\ dt \\ &ET\Delta t = \int_{t=0}^{\Delta t} et\ dt \\ &R\Delta t = \int_{t=0}^{\Delta t} r\ dt \\ &\Delta S = \int_{t=0}^{\Delta t} ds\ dt \end{aligned} \tag{2}$$

where P, ET, and R are the time-averaged precipitation rate (kg m^{-2} s^{-1}), the evapotranspiration rate (kg m^{-2} s^{-1}), and the runoff rate (kg m^{-2} s^{-1}) over the time period, respectively. ΔS is the change in water storage in the time period (kg m^{-2}). Note that the lower case letters (p, et, and r) denote instantaneous values and that the upper case letters (P, ET, and R) denote the time-averaged values over Δt.

As the length of the time period approaches infinity ($\Delta t \rightarrow \infty$), the terrestrial water storage is unable to be either replenished to infinity nor drained to negative infinity ($\infty > \Delta S > -\infty$). Thus, if the time period is sufficiently long (the third scientific hypothesis), then the terrestrial water storage term in Equation 2 can be neglected ($\lim\limits_{\Delta t \rightarrow \infty} (\Delta S/\Delta t) = 0$):

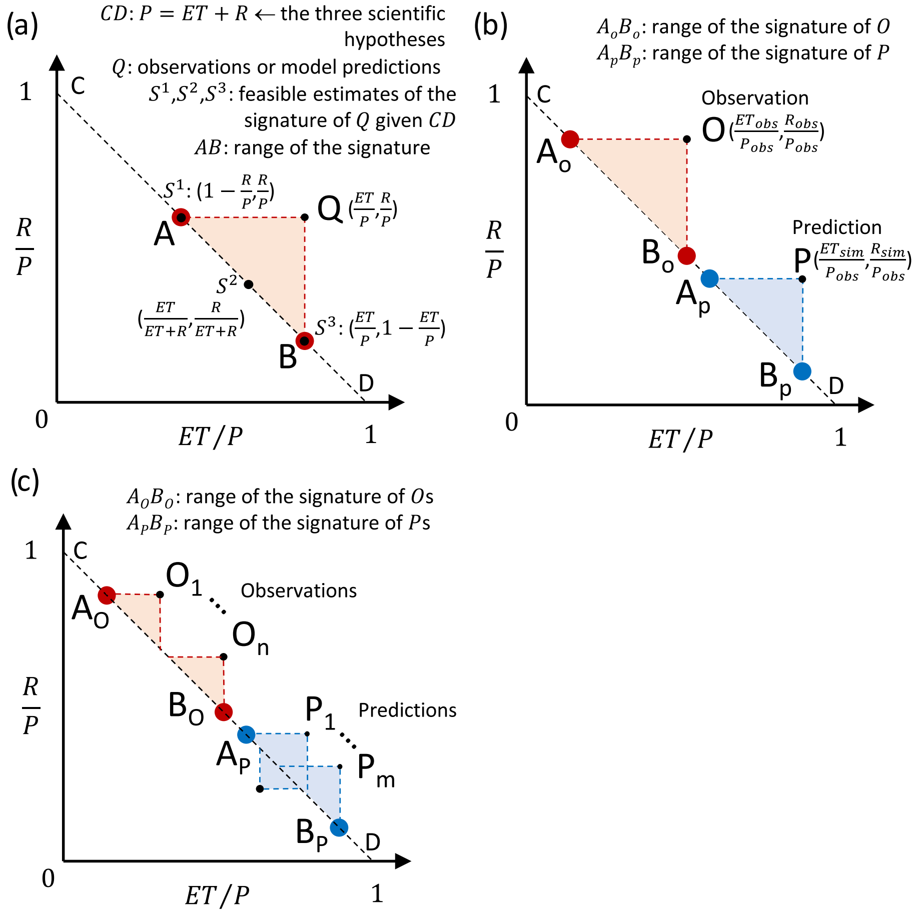
(a)
CD: $P = ET + R$ ← the three scientific hypotheses
Q: observations or model predictions
S^1,S^2,S^3: feasible estimates of the signature of Q given CD
AB: range of the signature
1
C
$\frac{R}{P}$
S^1: $(1-\frac{R}{P},\frac{R}{P})$
Q $(\frac{ET}{P},\frac{R}{P})$
A
S^2
$(\frac{ET}{ET+R},\frac{R}{ET+R})$
S^3: $(\frac{ET}{P},1-\frac{ET}{P})$
B
D
0
ET/P
1
(b)
A_oB_o: range of the signature of O
A_pB_p: range of the signature of P
1
C
Observation
O $(\frac{ET_{obs}}{P_{obs}},\frac{R_{obs}}{P_{obs}})$
A_o
$\frac{R}{P}$
B_o
Prediction
P $(\frac{ET_{sim}}{P_{obs}},\frac{R_{sim}}{P_{obs}})$
A_p
B_p
D
0
ET/P
1
(c)
A_OB_O: range of the signature of Os
A_PB_P: range of the signature of Ps
1
C
O_1
Observations
A_O
O_n
$\frac{R}{P}$
B_O
P_1
Predictions
A_P
P_m
B_P
D
0
ET/P
1

Figure 3. Falsification-oriented signature-based framework for evaluating land surface water balance. CD denotes the water balance line. As detailed in section 3.1, line CD is rigorously deduced from three scientific hypotheses: water mass conservation, no lateral flow of groundwater, and a sufficiently long time period. The signature in this framework is extracted from a set of precipitation (P), evapotranspiration (ET), and runoff (R) under the constraint of the three scientific hypotheses. The signatures must therefore lie on line CD, representing the long-term-averaged partitioning of precipitation between evapotranspiration and runoff. (a) Estimation the range of the signature of a point $Q(ET/P,R/P)$. Points S^1, S^2, and S^3 denote the feasible estimates of the signature by giving the three scientific hypotheses (line CD). They are obtained by combining P, ET, and R. Points A and B are set as the two ends of the three estimates. In this estimation, range AB is linearly proportional to the water budget imbalance ($\sqrt{2}\times|(P-ET-R)/P|$). (b) How to assert that a pair of observations and model predictions are inconsistent. A_oB_o and A_pB_p denote the ranges of the feasible signatures extracted from the observations (point O) and model predictions (point P), respectively. The consistency between the observations and the model predictions under the constraint of the three scientific hypotheses (line CD) is fully falsified if, and only if, the two signature ranges A_oB_o and A_pB_p do not overlap. (c) Falsifying rules for ensembles of observations and model predictions. The ensemble of model predictions is inconsistent with the observations if the two signature ranges A_OB_O and A_PB_P do not overlap.

$$1=\frac{ET}{P}+\frac{R}{P} \tag{3}$$

In Figure 3, Equation 3 is represented by the line CD (black dashed line) that connects the value of 1 on the x axis and the value of 1 on the y axis. Line CD is defined as the water balance line. If all three scientific hypotheses of the framework hold (i.e., water mass conservation, no horizontal flow, and a sufficiently long time period), then the point (ET/P, R/P) must lie on line CD.

3.2. Signature Extraction

As signatures are extracted based on these three scientific hypotheses, they must lie on the water balance line. The position on the line indicates the long-term partitioning of precipitation between evapotranspiration and runoff at the land surface.

Figure 3a shows how to extract the signature from point Q (ET/P, R/P). Point Q denotes a set of precipitation (P), evapotranspiration (ET), and runoff (R), which can be either observations or model predictions. Point Q may not lie on line CD. When point Q does not lie on line CD, its signature has a range. The range is estimated as follows. First, all feasible estimates of point Q's signature are obtained by combining the three scientific hypotheses (i.e., must lie on line CD) and two of the three water budget components (i.e., precipitation and evapotranspiration, evapotranspiration and runoff, and runoff and precipitation), which are points S^1, S^2, and S^3. Second, the upper and lower boundaries of these feasible estimates are obtained and denoted as points A and B. Note that range AB encompasses exactly all the feasible signatures extracted based on the three scientific hypotheses.

From Figure 3a, the precise positions of points A and B can be calculated as follows: (1) If point Q is above line CD ($P - ET - R < 0$), then points A and B are at ($1 - R/P$,R/P) and (ET/P,$1 - ET/P$), respectively; (2) if point Q is below line CD ($P - ET - R > 0$), then points A and B are at (ET/P,$1 - ET/P$) and ($1 - R/P$,R/P), respectively; and (3) if point Q is on line CD ($P - ET - R = 0$), then points A and B are the same as point Q (ET/P,R/P). As a result, the distance between points A and B is $\sqrt{2} \times |(P - ET - R)/P|$, which is linearly proportional to the water budget imbalance normalized by precipitation.

3.3. Falsifying Rules

Figure 3b shows the falsifying rules for a pair of model predictions and observations. Point P denotes the model predictions (ET_{sim}/P_{obs}, R_{sim}/P_{obs}), and point O denotes the observations (ET_{obs}/P_{obs}, R_{obs}/P_{obs}). A_pB_p and A_oB_o denote the range of feasible signatures extracted from the model predictions (point P) and the observations (point O), respectively. If the two ranges (A_pB_p and A_oB_o) do not overlap, then, given the three scientific hypotheses of the framework, the model predictions cannot be consistent with the observations. This is the fully falsified situation shown in Figure 2. There must be something wrong with the model predictions, the observations, and/or the scientific hypotheses for extracting signatures.

Figure 3c shows the falsifying rules for a pair of ensembles of model predictions and observations. Points P_1 to P_m denote the ensemble of model predictions, whereas points O_1 to O_n denote the ensemble of observations. For each model prediction and observation, feasible signatures are obtained following the steps described in section 3.2. Range A_PB_P denotes the range of the feasible signatures of all model predictions, and A_OB_O denotes the range of the feasible signatures of all observations. The two ensembles of model predictions and observations are asserted as inconsistent if A_PB_P and A_OB_O do not overlap.

4. Application Over the Continental United States

The framework is applied at each United States Geological Survey (USGS) eight-digit Hydrologic Unit (HUC8) basin over the CONUS. The observations used are described in section 4.1. The LSM used and its configurations are presented in section 4.2. The execution of the model simulations is described in section 4.3.

4.1. Multisource Observations Over the Continental USA

The precipitation data at a spatial resolution of 1/8th degree over the CONUS are from the North American Land Data Assimilation System (NLDAS). The NLDAS precipitation data (Xia et al., 2012; Xia et al., 2012) are derived from the gauged-only daily Climate Prediction Center (CPC) analysis data and adjusted for orographic effects based on the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climatology. The daily precipitation is disaggregated into hourly values (Xia, Mitchell, Ek, Cosgrove, et al., 2012; Xia, Mitchell, Ek, Sheffield, et al., 2012) based on the hourly weights from the State II Doppler radar derivations, the CPC MORPHing technique (CMORPH) satellite-based analyses, the CPC Hourly Precipitation Data Base (CPC HPD), and the North American Regional Reanalysis (NARR). The hourly data are used to drive the LSM.

The evapotranspiration data at a spatial resolution of 0.5° are derived by upscaling eddy covariance measurements of a global network (FLUXNET) using a multi-tree ensemble (MTE) method (Jung et al., 2009). The runoff data for each HUC8 are derived by the USGS. The evapotranspiration and runoff data have been widely used to evaluate NLDAS simulations over the CONUS (Xia et al., 2016). We upscaled the precipitation, evapotranspiration, and runoff observations at different spatial resolutions to the same USGS HUC8 basins.

Figure 1 shows the water budget imbalance in the three observational data sets, which is linearly proportional to the range of the signature of the observations. These water budgets are well balanced in the eastern United States, and the imbalance is within the range delineated by ±10% of precipitation. The balance reflects the fact that there are dense FLUXNET MTE training sites, rain gauges, and streamflow gauges in this region. There are no FLUXNET MTE training sites in eastern New Mexico, northwestern Texas, western Oklahoma, and western Kansas (Jung et al., 2009); the imbalance is significantly positive and about 25% of precipitation. In the mountainous areas of the West, the imbalance is significantly negative and even beyond −60% of precipitation, reflecting the sparsity of rain gauges and FLUXNET towers in the complex terrain and the significant measurement errors in the harsh environment.

4.2. Noah-MP LSM

Noah-MP (Niu et al., 2011; Yang et al., 2011) hosts multiple alternative options for several key process parameterizations and is therefore able to account for the known unknowns in these parameterizations. A 48-member ensemble is configured by combining four runoff options, three β-factor options (representing the control of soil moisture on transpiration), two stomatal conductance options, and two turbulence options. The four runoff options are divided into two groups. In the first group, there are two options with a groundwater component (Niu et al., 2005, 2007) as used in the Community Land Model (CLM) (Oleson et al., 2004). In the second group, the two options follow the Noah LSM (Chen & Dudhia, 2001) and the Biosphere Atmosphere Transfer System (BATS) (Dickinson et al., 1993), respectively. The three β-factor options are adopted from CLM (Oleson et al., 2004), Noah (Chen & Dudhia, 2001), and Simplified Simple Biosphere (SSiB) model (Xue et al., 1991). The two turbulence options (Brutsaert, 1982; Chen et al., 1997) are based on Monin-Obukhov similarity theory, which is commonly used in many LSMs. The two stomatal conductance options, the Jarvis (Chen et al., 1996) and Ball-Berry (Ball et al., 1987; Collatz et al., 1991, 1992) schemes, are used in second and third generation LSMs (Sellers et al., 1997), respectively. These parameterization options represent a spectrum of widely used LSMs, which Niu et al. (2011) and Yang et al. (2011) report dominate the hydrological simulations. Details of these options are described in Zheng et al. (2019).

All the Noah-MP parameters (Cuntz et al., 2016) use default values. The values have not been rigorously calibrated. With the falsification-oriented evaluation, parameters are not necessarily rigorously precalibrated. Miss-specified parameter values may result in inconsistencies between the model predictions and the observations, which can be detected and corrected in future experiments.

The observational data sets and models have been used for monitoring drought (https://ldas.gsfc.nasa.gov/nldas/drought-monitor) (Hao et al., 2016; Xia, Mitchell, Ek, Cosgrove, et al., 2012; Xia, Mitchell, Ek, Sheffield, et al., 2012) and predicting floods (https://water.noaa.gov) (Maidment, 2016; McEnery et al., 2005; Zheng et al., 2018) over the CONUS. They have been comprehensively evaluated with the verification-oriented approach. Cai, Yang, Xia, et al. (2014) and Cai, Yang, David et al. (2014) showed that Noah-MP outperforms the CLM, the Noah, and the variable infiltration capacity (VIC) models in replicating the observed soil moisture. Xia et al. (2016) found that Noah-MP outperforms Noah in reproducing the observed timing and magnitude of the monthly total runoff. The relative bias of simulated evapotranspiration from Noah-MP is 4% (Ma et al., 2017), and the modeled terrestrial water storage change anomaly is accurate (Ma et al., 2017; Xia et al., 2017).

4.3. Simulations

The atmospheric forcing, soil, and vegetation data at a spatial resolution of 1/8th degree are from NLDAS. The vegetation type data are the MODIS (Moderate Resolution Imaging Spectroradiometer) land cover type product classified using the International Geosphere-Biosphere Program scheme. The soil type data are derived from the State Soil Geographic database. For reference, the vegetation and soil types are shown geographically in Figure S1 in the supporting information.

The initial states of each simulation were obtained from a 102-year spin-up. The spin-up was performed in two steps. The models were run repeatedly 100 times over the year 1979 and then run through the 2 years from 1980 to 1981 with a time step of 15 min. The simulations for the following 30-year period from 1982 to 2011 were analyzed in this study.

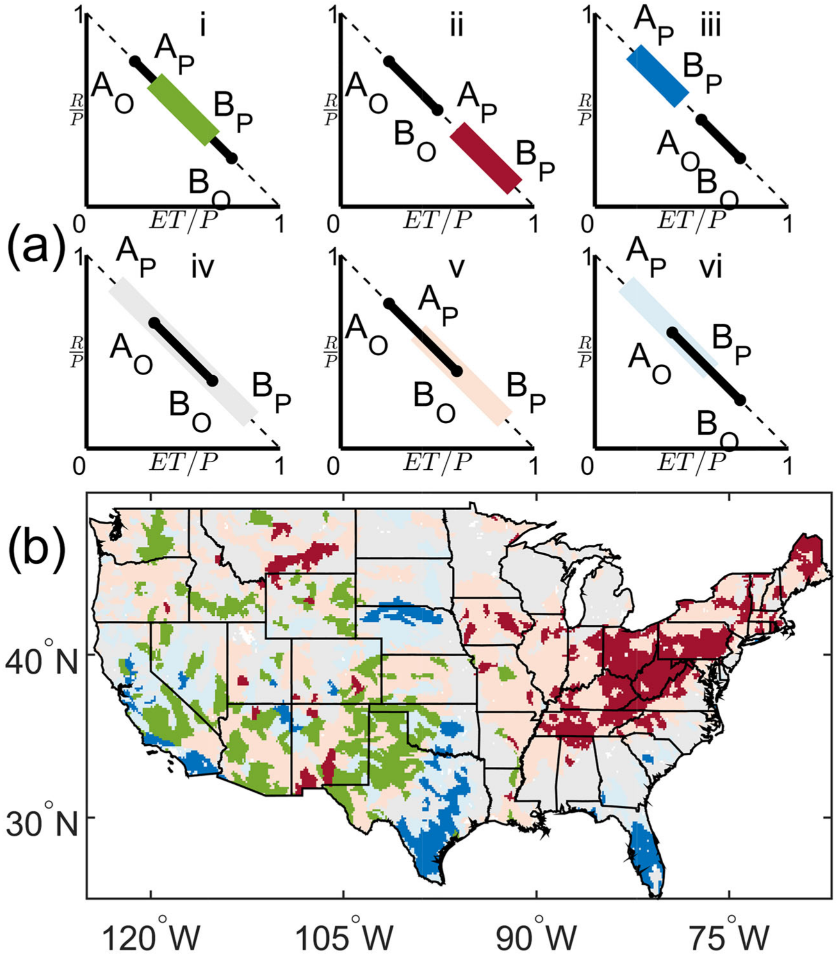

Figure 4. Evaluation of the Noah-MP ensemble over the continental United States. (a) A_OB_O denotes the range of the signature of observations, whereas A_PB_P denotes the range of the signature of the model predictions. The range A_OB_O is linearly proportional to the water budget imbalance, which is shown geographically in Figure 1. All six possible outcomes from the comparison of predictions and observations are shown in different colors: (i) not falsified with large observational uncertainty (green), (ii) fully falsified because all ensemble predictions overestimate the evapotranspiration fraction in precipitation (dark red), (iii) fully falsified because all ensemble predictions overestimate the runoff fraction in precipitation (dark blue), (iv) not falsified with a large modeling uncertainty (gray), (v) partially falsified because some predictions overestimate the evapotranspiration fraction, whereas none overestimates the runoff fraction (light red), and (vi) partially falsified because some overestimate runoff fraction, whereas none overestimates the evapotranspiration fraction (light blue). The geographical pattern of the evaluation results is shown in panel (b). The data for producing this figure can be found in the supporting information.

5. Results

Figure 4 shows the evaluation results over the CONUS. Specifically, emphasis is placed on three aspects: how the results help to guide future enhancements of the observations; insights for future improvements of the model as informed by the spatial patterns; and where the ensemble can and cannot outperform a single prediction.

Figure 4a shows the six types of outcomes from the evaluation. According to Figure 2 (section 2.2), they can be divided into three categories: (1) not falsified (types i and iv), (2) fully falsified (types ii and iii), or (3) partially falsified (types v and vi).

The nonfalsified category consists of two types: types i and iv. For type iv, the signature range of the predictions is larger than that of the observations. For type i, the signature range of the observations is larger than that of the predictions. Figure 4b shows that the type i nonfalsified situation occurs in western Texas, which is arid, and in the western United States, which is mountainous. As the signature range is linearly proportional to the water budget imbalance (section 3.2), the spatial patterns closely coincide with that of the water budget imbalance shown in Figure 1. The water budget imbalance (known unknowns in the observations) is too large to falsify any of the ensemble predictions. A top priority for the type i nonfalsified situation is to obtain physically consistent observations of precipitation, evapotranspiration, runoff, and lateral flows to close the water budget. Targeted observations should be pursued to address the measurement and processing errors.

The fully falsified situation could be characterized by either overestimating evapotranspiration (type ii) or runoff (type iii). Because all the predictions are inconsistent with the observations, the ensemble members share a common error. The ensemble prediction cannot outperform a single prediction in terms of matching the observations.

Figure 4b shows that all the ensemble predictions overestimate evapotranspiration and underestimate runoff (type ii) in the middle Ohio river basins and the Yellowstone river basin. The overestimation of evapotranspiration in the Ohio river basin may be attributable to model failure. Noah-MP introduced an explicit canopy layer, enabling the modeling of canopy interception and evaporation. It is a major structural augment over the Noah LSM. The Ohio river basins are covered by deciduous broadleaf forest (Figure S1b), the canopy of which is highly capable of intercepting precipitation. Canopy interception and evaporation are strong over the middle Ohio river basins and significantly different across the Noah-MP ensemble (Zheng et al., 2019). The canopy interception and loss may be too strong, leading to an overall overestimation of evapotranspiration and an underestimation of runoff. The Noah-MP parameterizations and parameters for canopy interception and loss should therefore be scrutinized.

The overestimation of evapotranspiration in the Yellowstone river basin (Wyoming) may be linked to groundwater recharge driven by the geothermal plumbing. The Yellowstone river basin exhibits complex geothermal activity and has the largest collection of geysers on Earth. Groundwater seeps down into the geothermal plumbing to be heated by the Yellowstone megavolcano and then flows into rivers and lakes through hot springs, geysers, and mud pots. While increasing runoff, the geothermal plumbing process drives groundwater recharge from soil moisture and lowers evapotranspiration. The Noah-MP does not represent these processes and therefore overestimates evapotranspiration and underestimates runoff.

Figure 4b shows that all the ensemble predictions underestimate evapotranspiration and overestimate runoff (type iii) in central Florida, eastern Texas, the Niobrara river basin (Nebraska), and the Salton Sea river basin (southern California). Interestingly, all these basins are in groundwater discharge areas. The majority of the water in the Niobrara river comes from groundwater seepage from the High Plains, or Ogallala, aquifer. Central Florida is occupied by the lower Floridan aquifer, the water of which comes from the upper Floridan aquifer in Alabama and Georgia and may also be intruded by seawater. The Texas basins are precisely downstream (east) of the Balcones fault zones. The Salton Sea in southern California is one of the world's largest inland lakes. The surface elevation in the Salton Sea is about 70 m below sea level, allowing the inflow of water from the surrounding mountainous.

There are several possible reasons for the falsification in these basins. First, groundwater discharge provides additional soil moisture for evapotranspiration, which can be captured by the evapotranspiration observations but not by the Noah-MP model. All the Noah-MP ensemble predictions therefore underestimate evapotranspiration. Second, all these basins coincide with the sand soil type (Figure S1). Noah-MP estimates the permeability to be the same as sand. However, the estimated permeability is probably too low for the permeable surface in the groundwater discharge zones (e.g., macropores) and causes an overestimation of the runoff. For instance, the fully falsified Texas basins are rich in vertisols. Vertisols can have deep and wide cracks, which are preferential pathways for water flow (Kurtzman et al., 2016). Third, groundwater discharge zones often exhibit heavy human activity. Humans extract water from groundwater and rivers into the soil moisture pool for agriculture. These activities increase evapotranspiration and decrease runoff, which is not considered in Noah-MP.

Despite the challenges of pinpointing the causes of the fully falsified situation, Figure 4 shows that the falsification-oriented evaluation is powerful in identifying areas of modeling/observational challenges. These results can stimulate the development of new scientific hypotheses and observations. Unless these missing processes (i.e., unknown unknowns) have been correctly represented by the model and measured by the observations, ensemble predictions are not very helpful because all the ensemble members have been falsified. In addition, verification-oriented calibration is not expected to be a robust solution for improving the simulations for these basins as a result of as yet unknown processes.

Interestingly, these nonfalsified and fully falsified basins have a remarkable overlap with the locations of Critical Zone Observatories (CZOs) (Dybas, 2013) (Figure S4). The Southern Sierra (Holbrook et al., 2014; Visser et al., 2019), Reynolds Creek (Seyfried et al., 2018), and Catalina-Jemez (McIntosh et al., 2017) CZOs are in basins where the water budget imbalance is too large to falsify any model predictions (type i). The Clear Creek, Susquehanna Shale Hills (Jin et al., 2011; Liu et al., 2020), and Christina river basin CZOs are in basins where all the model predictions are falsified by overestimating evapotranspiration and underestimating runoff (type ii). The evaluation results also highlight the need for strengthening observatories in Texas and Florida, where all the Noah-MP predictions are falsified by overestimating the runoff and underestimating the evapotranspiration (type iii). The Texas Water Observatory infrastructures have already been installed over these areas. New theoretical advancements and observational data (Fan, 2015) from these observatories are expected to resolve the water budget imbalance and the inconsistencies between the model predictions and the observations.

The partially falsified situations (types v and vi) hint at the importance of model intercomparison and statistical postprocessing. The ensemble predictions are distinguishable in terms of matching the observations. By intercomparing the poorly performing predictions and those that performed well, known unknowns can be solved, and known knowns grow. The range of the signature of predictions can thus be reduced, which is important for future falsification-oriented evaluations (Figure 2). Statistical methods are also beneficial in these situations, and an ensemble can produce better predictions for various applications by selecting and weighting the ensemble members.

6. Discussion

Falsification- and verification-oriented evaluations have the same target: guiding the development of models and the enhancement of observations. The difference is that falsification exposes “mistakes” in, or inconsistencies between, the model predictions and the observations. The “mistakes” should be carefully identified, reported, analyzed, and fixed. This study identified and reported such “mistakes” in the modeled and observed climatology over the CONUS with a newly developed framework.

The proposed framework assumes that the time period is long enough to neglect the change in terrestrial water storage. In this study, limited by the observations, the period spans 30 years (1982–2011). This time period should be sufficiently long because the terrestrial water storage is characterized by generally stable annual cycles (Lorenz et al., 2014). A 30-year time period length is long enough for even decadal drought events. During the Millennium Drought from 2001 to 2009 in southeast Australia (van Dijk et al., 2013), terrestrial water storage decreased by roughly 60 mm, while the annual average precipitation during the drought was approximately 450 mm. The ratio of the change in terrestrial water storage to the product of the annual average precipitation and 30 years is $\leq 0.5\%$, which can be neglected. Using satellite observations from the Gravity Recovery and Climate Experiment (GRACE) and the Noah-MP model predictions over the CONUS, Figures S2 and S3 also show that the terrestrial water storage term is neglectable compared with the precipitation term in Equation 2. We therefore argue that a time period of 30 years is generally sufficient to accumulate enough amount of precipitation to neglect the change in terrestrial water storage in Equation 2.

However, because the change in terrestrial water storage is neglected, this framework is limited to evaluating the climatology (i.e., the long-term-averaged partitioning of precipitation between evapotranspiration and runoff). More advanced signatures should be developed to evaluate hydrometeorological variations (e.g., droughts and floods). These signatures should consider both the terrestrial water storage itself and its change to fully characterize the trajectory of the dynamics of the terrestrial water system in phase space (i.e., state and change). We also expect that the evaluation is problem-oriented (e.g., frequency, magnitude, and the onset of extreme events) and differs in time scales (e.g., daily, monthly, or yearly). This warrants further investigations.

Human activities do not affect the deduction of the framework described in section 3. However, human activities can complicate the validity of the three scientific hypotheses and, thus, require careful attention when interpreting a falsified result. First, the third scientific hypothesis of the framework, which states that 30 years is sufficiently long to neglect the change in terrestrial water storage, is very likely to hold. From Figure S2b, the human-induced change in the terrestrial water storage is unlikely to be comparable with that caused by drought. The change in terrestrial water storage can be safely neglected in comparison with the long-term accumulated amount of precipitation. Falsified results are therefore unlikely to be attributable to the failure of this assumption. Second, groundwater withdrawal can drive horizontal replenishment from the surrounding areas (the second scientific hypothesis of the framework). The lateral flow of groundwater can increase or decrease the water budget imbalance, depending on the existing imbalance in the observations caused by errors, and may play a part in the falsified results. Third, the use of water by humans can redistribute the water locally among different components of the water budget. Agricultural water use may reduce runoff and increase evapotranspiration as compensation. If the redistribution is sufficiently significant, then the model predictions without considering the human use of water should be inconsistent with the observations. In a falsified result, the failure in modeling the human use of water is often accounted for.

The consideration of known unknowns can be improved further. This study only considered the water budget imbalance resulting from all the observable terms as a whole (Beven & Westerberg, 2011). In fact, other known unknowns exist in each of the observational terms. Sun et al. (2018) provided a comprehensive overview of 30 precipitation products and concluded that their reliability is mainly limited by the number and spatial coverage of surface stations, the satellite algorithms, and the data assimilation models. Wang and Dickinson (2012) provided a comprehensive review of evapotranspiration observations and showed that the unknowns in the underlying theory (i.e., the Monin-Obukhov similarity theory), the spatial coverage of surface stations, satellite algorithms, and data postprocessing techniques (e.g., gap filling) have significant impacts. Di Baldassarre and Montanari (2009) argued that the unknowns in large-scale runoff observations are also far from negligible. The terrestrial water cycle is also influenced by human activities, which may not have been reflected in the evapotranspiration and runoff observations. However, compared with the spread among multiple observations of precipitation (Xia et al., 2016) and evapotranspiration (Long et al., 2014) over the CONUS, the water budget imbalance among them (Figure 1) is much more significant over most areas. The neglect of the known unknowns within each observation should therefore be generally valid, except for those areas with a neglectable water budget imbalance. The method for signature extraction (section 3.2) is extendable to include these known unknowns.

7. Conclusions

There are always known knowns, known unknowns, and unknown unknowns in the modeling and observation of land surface processes. Large-scale land surface modeling provides a baseline of “modeling everywhere as a learning process” (Beven, 2007; Beven & Alcock, 2012; Beven & Cloke, 2012; Wood et al., 2011) and is vital to advance Earth system modeling (Archfield et al., 2015; Bierkens, 2015; Clark et al., 2015). This study proposes the falsification-oriented signature-based evaluation. This method recognizes the asymmetry between verification and falsification and “learns from mistakes” through a “try-fail-refine” strategy. With this approach, not only models but also observations and evaluation methods can be evaluated simultaneously.

Verification is valuable, but verification-oriented evaluation results should be interpreted with the caution of overconfidence. Verification-oriented evaluation often implicitly assumes that if the model predictions are consistent with every known observation in every known aspect, then everything is correct. This assumption is conditional on the negligibility of unknown unknowns. Only if unknown unknowns are neglectable, the likelihood of a model being true can be estimated. In the practice of verification-oriented evaluation, unknown unknowns are often lumped into known unknowns (e.g., a lumped a priori error distribution). Overconfidence in the evaluation results may hinder scientific explorations into unknown unknowns for new observations and theories.

The test of the consistency between model predictions and observations is asymmetrical between verification- and falsification-oriented approaches. One single inconsistency always indicates that there must be something wrong. Inconsistency is a solid indicator of errors in any characteristic (i.e., known unknowns and unknown unknowns) and from any source (e.g., observations, parameterizations, parameters, evaluation methods, and coding bugs). The solidity of the inconsistency as an indicator of “mistakes” is important for making concrete progress in both model developments and observation enhancements. Falsification-oriented evaluation values and learns from these inconsistencies.

The proposed framework is shown to be powerful in revealing the areas with modeling and observational challenges. In the mountainous and arid areas of the western CONUS, the water budget imbalance among observations is too large to falsify a model prediction. The imbalance results from the unobserved lateral flow and the observational errors in precipitation, evapotranspiration, and runoff. In these areas, enhancements of the observations are more urgently needed than model improvements. The imbalance in the eastern United States is much smaller than that in the West. The underestimation of evapotranspiration and overestimation of runoff occur simultaneously in the HUC8 basins of central Florida on the lower Floridan aquifer, the eastern Texas basins just downstream of the Balcones fault zones, the Niobrara river basin at the north tip of the Ogallala aquifer, and the Salton Sea river basin. The middle Ohio river basins and the Yellowstone river basin are falsified by overestimating evapotranspiration and underestimating runoff. The falsified results may be attributable to the failure of Noah-MP in representing region-specific processes and human activities. A top priority for these fully falsified situations is to develop new scientific hypotheses and observations to close the gap between the model predictions and the observations.

Interestingly, a substantial portion of the CZOs is located in the regions with nonfalsified or fully falsified evaluation results. These established CZOs include Southern Sierra, Reynolds Creek, Catalina-Jemez, Clear Creek, Susquehanna Shale Hills, and Christina River Basin. New theoretical and observational advancements from these observatories are expected to resolve the water budget imbalance and the inconsistencies between model predictions and observations. Our results also highlight the need for strengthening the Texas Water Observatory and for extending observatories in Florida.

Data Availability Statement

The NLDAS static data and meteorological forcing data were obtained from the Goddard Earth Sciences Data and Information Services Center (https://disc.sci.gsfc.nasa.gov/uui/datasets?keywords=NLDAS). The FLUXNET MTE evapotranspiration data were obtained from the Max Planck Institute for Biogeochemistry (https://www.bgc-jena.mpg.de/geodb/projects/Home.php). The USGS HUC8 runoff data were downloaded from the USGS WaterWatch website (https://waterwatch.usgs.gov/?id=romap3). The model outputs were generated from the Noah-MP version 3.6, which can be downloaded from the website of the Research and Applications Laboratory at the National Center for Atmospheric Research (https://ral.ucar.edu/solutions/products/noah-multiparameterization-land-surface-model-noah-mp-lsm). The evaluation results at each HUC8 basins can be found in the supporting information.

Acknowledgments

We thank the reviewers and editors for their detailed and insightful comments. We thank Dr. Peili Wu at the Met Office Hadley Center for his generous help with structuring and writing this paper. This work is jointly supported by the National Key Research and Development Program of China grants 2018YFA0606004 and 2016YFA0600403, the National Natural Science Foundation of China grants 41605062 and 41375088, and the Science and Technology Project of State Grid Corporation of China.

References

Adam, J. C., Clark, E. A., Lettenmaier, D. P., & Wood, E. F. (2006). Correction of global precipitation products for orographic effects. *Journal of Climate*, *19*(1), 15–38. https://doi.org/10.1175/JCLI3604.1

Archfield, S. A., Clark, M. P., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., et al. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, *51*, 10,078–10,091. https://doi.org/10.1002/2015WR017498

Baker, V. R. (2017). Debates-hypothesis testing in hydrology: Pursuing certainty versus pursuing uberty. *Water Resources Research*, *53*, 1770–1778. https://doi.org/10.1002/2016WR020078

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. In J. Biggins (Ed.), *Progress in Photosynthesis Research* (pp. 221–224). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-0519-6_48

Beven, K. J. (1993). Prophecy, reality and uncertainty in distributed hydrological modelling. *Advances in Water Resources*, *16*(1), 41–51. https://doi.org/10.1016/0309-1708(93)90028-E

Beven, K. J. (2001). On hypothesis testing in hydrology. *Hydrological Processes*, *15*(9), 1655–1657. https://doi.org/10.1002/hyp.436

Beven, K. J. (2004). Does an interagency meeting in Washington imply uncertainty? *Hydrological Processes*, *18*(9), 1747–1750. https://doi.org/10.1002/hyp.5573

Beven, K. J. (2007). Towards integrated environmental models of everywhere: Uncertainty, data and modelling as a learning process. *Hydrology and Earth System Sciences*, *11*(1), 460–467. https://doi.org/10.5194/hess-11-460-2007

Beven, K. J. (2009). *Environmental Modelling: An Uncertain Future?* London: Routledge.

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *Wiley Interdisciplinary Reviews Water*, *5*(3), e1278. https://doi.org/10.1002/wat2.1278

Beven, K. J., & Alcock, R. E. (2012). Modelling everything everywhere: A new approach to decision-making for water management under uncertainty. *Freshwater Biology*, *57*(SUPPL. 1), 124–132. https://doi.org/10.1111/j.1365-2427.2011.02592.x

Beven, K. J., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K. J., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes*, *28*(24), 5897–5918. https://doi.org/10.1002/hyp.10082

Beven, K. J., & Cloke, H. L. (2012). Comment on "Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al. *Water Resources Research*, *48*, W01801. https://doi.org/10.1029/2011WR010982

Beven, K. J., Smith, P. J., & Wood, A. (2011). On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, *15*(10), 3123–3133. https://doi.org/10.5194/hess-15-3123-2011

Beven, K. J., & Westerberg, I. (2011). On red herrings and real herrings: Disinformation and information in hydrological inference. *Hydrological Processes*, *25*(10), 1676–1680. https://doi.org/10.1002/hyp.7963

Bierkens, M. F. P. (2015). Global hydrology 2015: State, trends, and directions. *Water Resources Research*, *51*, 4923–4947. https://doi.org/10.1002/2015WR017173

Binley, A. M., Beven, K. J., Calver, A., & Watts, L. G. (1991). Changing responses in hydrology: Assessing the uncertainty in physically based model predictions. *Water Resources Research*, *27*(6), 1253–1261. https://doi.org/10.1029/91WR00130

Blöschl, G. (2017). Debates-hypothesis testing in hydrology: Introduction. *Water Resources Research*, *53*, 1767–1769. https://doi.org/10.1002/2017WR020584

Brutsaert, W. (1982). *Evaporation Into the Atmosphere: Theory, History, and Applications*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-1497-6

Cai, X., Yang, Z.-L., David, C. H., Niu, G.-Y., & Rodell, M. (2014). Hydrological evaluation of the Noah-MP land surface model for the Mississippi River Basin. *Journal of Geophysical Research: Atmospheres*, *119*, 23–38. https://doi.org/10.1002/2013JD020792

Cai, X., Yang, Z.-L., Xia, Y., Huang, M., Wei, H., Leung, L. R., & Ek, M. B. (2014). Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS test bed. *Journal of Geophysical Research: Atmospheres*, *119*, 13,751–13,770. https://doi.org/10.1002/2014JD022113

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. https://doi.org/10.1175/1520-0493(2001)129<0587:caalsh>2.0.co;2

Chen, F., Janjić, Z., & Mitchell, K. (1997). Impact of atmospheric surface-layer parameterizations in the new land-surface scheme of the NCEP mesoscale Eta model. *Boundary-Layer Meteorology*, *85*(3), 391–421. https://doi.org/10.1023/A:1000531001463

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H.-L., Koren, V., et al. (1996). 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

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., et al. (2015). Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research*, *51*, 5929–5956. https://doi.org/10.1002/2015WR017096

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*, W09301. https://doi.org/10.1029/2010WR009827

Collatz, G. J., Ball, J. T., Grivet, C., & Berry, J. A. (1991). Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: A model that includes a laminar boundary layer. *Agricultural and Forest Meteorology*, *54*(2), 107–136. https://doi.org/10.1016/0168-1923(91)90002-8

Collatz, G. J., Ribas-Carbo, M., & Berry, J. A. (1992). Coupled photosynthesis-stomatal conductance model for leaves of C4 plants. *Functional Plant Biology*, *19*(5), 519–538. https://doi.org/10.1071/pp9920519

Cuntz, M., Mai, J., Samaniego, L., Clark, M. P., Wulfmeyer, V., Branch, O., et al. (2016). The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model. *Journal of Geophysical Research: Atmospheres*, *121*, 10,676–10,700. https://doi.org/10.1002/2016JD025097

Di Baldassarre, G., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. *Hydrology and Earth System Sciences*, *13*(6), 913–921. https://doi.org/10.5194/hess-13-913-2009

Dickinson, R. E., Henderson-Sellers, A., & Kennedy, P. J. (1993). *Biosphere-Atmosphere Transfer Scheme (BATS) Version 1e as Coupled to the NCAR Community Climate Model*. Boulder, CO: University Corporation for Atmospheric Research. https://doi.org/10.5065/D67W6959

Dybas, C. (2013). *Discoveries in the Critical Zone (CZO): Where Life Meets Rock (No. nsf13112)*. Alexandria, VA: National Science Foundation. Retrieved from https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf13112

Eagleson, P. S. (1978). Climate, soil, and vegetation: 1. Introduction to water balance dynamics. *Water Resources Research*, *14*(5), 705–712. https://doi.org/10.1029/WR014i005p00705

Fan, Y. (2015). Groundwater in the Earth's critical zone: Relevance to large-scale patterns and processes. *Water Resources Research*, *51*, 3052–3069. https://doi.org/10.1002/2015WR017037

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Hao, Z., Hong, Y., Xia, Y., Singh, V. P., Hao, F., & Cheng, H. (2016). Probabilistic drought characterization in the categorical form using ordinal regression. *Journal of Hydrology*, *535*, 331–339. https://doi.org/10.1016/j.jhydrol.2016.01.074

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D. (2018). An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology*, *556*, 1205–1219. https://doi.org/10.1016/j.jhydrol.2017.03.008

Hill, M. C., Foglia, L., Christensen, S., Rakovec, O., & Borgonovo, E. (2017). Model validation: Testing models using data and sensitivity analysis. In J. H. Cushman, & D. M. Tartakovsky (Eds.), *The Handbook of Groundwater Engineering* (3rd ed. pp. 597–624). Boca Raton, Florida: CPC Press. https://doi.org/10.1201/9781315371801

Holbrook, W. S., Riebe, C. S., Elwaseif, M., Hayes, L., Basler-Reeder, K., Harry, L., et al. (2014). Geophysical constraints on deep weathering and water storage potential in the Southern Sierra Critical Zone Observatory. *Earth Surface Processes and Landforms*, *39*(3), 366–380. https://doi.org/10.1002/esp.3502

Hubble, E. (2013). *The Realm of the Nebulae*. New Haven, Connecticut: Yale University Press.

Jin, L., Andrews, D. M., Holmes, G. H., Lin, H., & Brantley, S. L. (2011). Opening the "black box": Water chemistry reveals hydrological controls on weathering in the Susquehanna Shale Hills Critical Zone Observatory. *Vadose Zone Journal*, *10*(3), 928–942. https://doi.org/10.2136/vzj2010.0133

Jung, M., Reichstein, M., & Bondeau, A. (2009). Towards global empirical upscaling of FLUXNET eddy covariance observations: Validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, *6*(10), 2001–2013. https://doi.org/10.5194/bg-6-2001-2009

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., & Westerberg, I. K. (2013). Disinformative data in large-scale hydrological modelling. *Hydrology and Earth System Sciences*, *17*(7), 2845–2857. https://doi.org/10.5194/hess-17-2845-2013

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*, W03S04. https://doi.org/10.1029/2005WR004362

Kurtzman, D., Baram, S., & Dahan, O. (2016). Soil-aquifer phenomena affecting groundwater under vertisols: A review. *Hydrology and Earth System Sciences*, *20*(1), 1–12. https://doi.org/10.5194/hess-20-1-2016

Linde, N. (2014). Falsification and corroboration of conceptual hydrological models using geophysical data. *Wiley Interdisciplinary Reviews Water*, *1*(2), 151–171. https://doi.org/10.1002/wat2.1011

Liu, H., Yu, Y., Zhao, W., Guo, L., Liu, J., & Yang, Q. (2020). Inferring subsurface preferential flow features from a wavelet analysis of hydrological signals in the Shale Hills catchment. *Water Resources Research*, *56*, e2019WR026668. https://doi.org/10.1029/2019WR026668

Long, D., Longuevergne, L., & Scanlon, B. R. (2014). Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites. *Water Resources Research*, *50*, 1131–1151. https://doi.org/10.1002/2013WR014581

Lorenz, C., Kunstmann, H., Devaraju, B., Tourian, M. J., Sneeuw, N., & Riegger, J. (2014). Large-scale runoff from landmasses: A global assessment of the closure of the hydrological and atmospheric water balances. *Journal of Hydrometeorology*, *15*(6), 2111–2139. https://doi.org/10.1175/JHM-D-13-0157.1

Ma, N., Niu, G.-Y., Xia, Y., Cai, X., Zhang, Y., Ma, Y., & Fang, Y. (2017). A systematic evaluation of Noah-MP in simulating land-atmosphere energy, water, and carbon exchanges over the continental United States. *Journal of Geophysical Research: Atmospheres*, *122*, 12,245–12,268. https://doi.org/10.1002/2017JD027597

Maidment, D. R. (2016). Conceptual framework for the National Flood Interoperability Experiment. *Journal of the American Water Resources Association*, *53*(2), 245–257. https://doi.org/10.1111/1752-1688.12474

McEnery, J., Ingram, J., Duan, Q., Adams, T., & Anderson, L. (2005). NOAA's advanced hydrologic prediction service. *Bulletin of the American Meteorological Society*, *86*(3), 375–386. https://doi.org/10.1175/BAMS-86-3-375

McIntosh, J. C., Schaumberg, C., Perdrial, J., Harpold, A., Vázquez-Ortega, A., Rasmussen, C., et al. (2017). Geochemical evolution of the Critical Zone across variable time scales informs concentration-discharge relationships: Jemez River Basin Critical Zone Observatory. *Water Resources Research*, *53*, 4169–4196. https://doi.org/10.1002/2016WR019712

McKnight, D. M. (2017). Debates-hypothesis testing in hydrology: A view from the field: The value of hydrologic hypotheses in designing field studies and interpreting the results to advance hydrology. *Water Resources Research*, *53*, 1779–1783. https://doi.org/10.1002/2016WR020050

Neuweiler, I., & Helmig, R. (2017). Debates-hypothesis testing in hydrology: A subsurface perspective. *Water Resources Research*, *53*, 1784–1791. https://doi.org/10.1002/2016WR020047

Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., & Gulden, L. E. (2005). A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *Journal of Geophysical Research*, *110*, D21106. https://doi.org/10.1029/2005JD006111

Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., Gulden, L. E., & Su, H. (2007). Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *Journal of Geophysical Research*, *112*, D07103. https://doi.org/10.1029/2006JD007522

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (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

Oleson, K. W., Dai, Y., Bonan, G. B., Bosilovich, M., Dirmeyer, P. A., & Hoffman, F. M. (2004). *Technical Description of the Community Land Model (CLM)*. Boulder, CO: University Corporation for Atmospheric Research. https://doi.org/10.5065/D6N877R0

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2012). Multisource estimation of long-term terrestrial water budget for major global river basins. *Journal of Climate*, *25*(9), 3191–3206. https://doi.org/10.1175/JCLI-D-11-00300.1

Pfister, L., & Kirchner, J. W. (2017). Debates-hypothesis testing in hydrology: Theory and practice. *Water Resources Research*, *53*, 1792–1798. https://doi.org/10.1002/2016WR020116

Popper, K. R. (2002). *Conjectures and Refutations: The Growth of Scientific Knowledge* (2nd ed.). London: Routledge.

Refsgaard, J. C., & Henriksen, H. J. (2004). Modelling guidelines––Terminology and guiding principles. *Advances in Water Resources*, *27*(1), 71–82. https://doi.org/10.1016/j.advwatres.2003.08.006

Reggiani, P., & Schellekens, J. (2003). Modelling of hydrological responses: The representative elementary watershed approach as an alternative blueprint for watershed modelling. *Hydrological Processes*, *17*(18), 3785–3789. https://doi.org/10.1002/hyp.5167

Reggiani, P., Sivapalan, M., & Hassanizadeh, S. M. (2000). Conservation equations governing hillslope responses: Exploring the physical basis of water balance. *Water Resources Research*, *36*(7), 1845–1863. https://doi.org/10.1029/2000WR900066

Reggiani, P., Sivapalan, M., Hassanizadeh, S. M., & Gray, W. G. (2001). Coupled equations for mass and momentum balance in a stream network: Theoretical derivation and computational experiments. *Proceedings of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, *457*(2005), 157–189. https://doi.org/10.1098/rspa.2000.0661

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2011). Reconciling the global terrestrial water budget using satellite remote sensing. *Remote Sensing of Environment*, *115*(8), 1850–1865. https://doi.org/10.1016/j.rse.2011.03.009

Sellers, P. J., Dickinson, R. E., Randall, D. A., Betts, A. K., Hall, F. G., Berry, J. A., et al. (1997). Modeling the exchanges of energy, water, and carbon between continents and the atmosphere. *Science*, *275*(5299), 502–509. https://doi.org/10.1126/science.275.5299.502

Seyfried, M., Lohse, K. A., Marks, D. G., Flerchinger, G. N., Pierson, F., & Holbrook, W. S. (2018). Reynolds Creek experimental watershed and Critical Zone Observatory. *Vadose Zone Journal*, *17*(1), 1–20. https://doi.org/10.2136/vzj2018.07.0129

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F., & McCabe, M. F. (2009). Closing the terrestrial water budget from satellite remote sensing. *Geophysical Research Letters*, *36*, L07403. https://doi.org/10.1029/2009GL037338

Sherwood, S. (2011). Science controversies past and present. *Physics Today*, *64*(10), 39–44. https://doi.org/10.1063/PT.3.1295

Sivapalan, M., Blöschl, G., Zhang, L., & Vertessy, R. (2003). Downward approach to hydrological prediction. *Hydrological Processes*, *17*(11), 2101–2111. https://doi.org/10.1002/hyp.1425

Son, K., & Sivapalan, M. (2007). Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resources Research*, *43*, W01415. https://doi.org/10.1029/2006WR005032

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K.-L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, *56*, 79–107. https://doi.org/10.1002/2017RG000574

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review*, *38*(1), 55. https://doi.org/10.2307/210739 van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., et al. (2013). The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, *49*, 1040–1057. https://doi.org/10.1002/wrcr.20123

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., et al. (2017). Pedotransfer functions in earth system science: Challenges and perspectives. *Reviews of Geophysics*, *55*, 1199–1256. https://doi.org/10.1002/2017RG000581

Visser, A., Thaw, M., Deinhart, A., Bibby, R., Safeeq, M., Conklin, M., et al. (2019). Cosmogenic isotopes unravel the hydrochronology and water storage dynamics of the southern sierra critical zone. *Water Resources Research*, *55*, 1429–1450. https://doi.org/10.1029/2018WR023665

Wagener, T., & Montanari, A. (2011). Convergence of approaches toward reducing uncertainty in predictions in ungauged basins. *Water Resources Research*, *47*, W06301. https://doi.org/10.1029/2010WR009469

Wang, K., & Dickinson, R. E. (2012). A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability. *Reviews of Geophysics*, *50*, RG2005. https://doi.org/10.1029/2011RG000373

Wilby, R. L., Clifford, N. J., De Luca, P., Harrigan, S., Hillier, J. K., Hodgkins, R., et al. (2017). The 'dirty dozen' of freshwater science: Detecting then reconciling hydrological data biases and errors. *Wiley Interdisciplinary Reviews Water*, *4*(3), e1209. https://doi.org/10.1002/wat2.1209

Wilm, H. G., Thornthwaite, C. W., Colman, E. A., Cummings, N. W., Croft, A. R., Gisborne, H. T., et al. (1944). Report of the committee on transpiration and evaporation, 1943–44. *Transactions of the American Geophysical Union*, *25*(5), 683. https://doi.org/10.1029/TR025i005p00683

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*, W05301. https://doi.org/10.1029/2010WR010090

Xia, Y., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Brewer, M., et al. (2016). 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., Mitchell, K., Ek, M., Cosgrove, B. A., Sheffield, J., Luo, L., et al. (2012). 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. A., Wood, E. F., et al. (2012). 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*, *117*, D03109. https://doi.org/10.1029/2011JD016048

Xia, Y., Mocko, D., Huang, M., Li, B., Rodell, M., Mitchell, K. E., et al. (2017). Comparison and assessment of three advanced land surface models in simulating terrestrial water storage components over the United States. *Journal of Hydrometeorology*, *18*(3), 625–649. https://doi.org/10.1175/JHM-D-16-0112.1

Xue, Y., Sellers, P. J., Kinter, J. L., & Shukla, J. (1991). A simplified biosphere model for global climate studies. *Journal of Climate*, *4*(3), 345–364. https://doi.org/10.1175/1520-0442(1991)004<0345:asbmfg>2.0.co;2

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (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

Zheng, H., Yang, Z.-L., Lin, P., Wei, J., Wu, W.-Y., Li, L., et al. (2019). On the sensitivity of the precipitation partitioning into evapotranspiration and runoff in land surface parameterizations. *Water Resources Research*, *55*, 95–111. https://doi.org/10.1029/2017WR022236

Zheng, X., Maidment, D. R., Tarboton, D. G., Liu, Y. Y., & Passalacqua, P. (2018). GeoFlood: Large-scale flood inundation mapping based on high-resolution terrain analysis. *Water Resources Research*, *54*(12), 10,013–10,033. https://doi.org/10.1029/2018WR023457