Multiscale and Multivariate Evaluation of Water Fluxes and States over European River Basins

OLDRICH RAKOVEC, ROHINI KUMAR, JULIANE MAI, MATTHIAS CUNTZ, STEPHAN THOBER, MATTHIAS ZINK, SABINE ATTINGER, DAVID SCHÄFER, MARTIN SCHRÖN, AND LUIS SAMANIEGO

Helmholtz Centre for Environmental Research–UFZ, Leipzig, Germany

ABSTRACT

Accurately predicting regional-scale water fluxes and states remains a challenging task in contemporary hydrology. Coping with this grand challenge requires, among other things, a model that makes reliable predictions across scales, locations, and variables other than those used for parameter estimation. In this study, the mesoscale hydrologic model (mHM) parameterized with the multiscale regionalization technique is comprehensively tested across 400 European river basins. The model fluxes and states, constrained using the observed streamflow, are evaluated against gridded evapotranspiration, soil moisture, and total water storage anomalies, as well as local-scale eddy covariance observations. This multiscale verification is carried out in a seamless manner at the native resolutions of available datasets, varying from 0.5 to 100 km. Results of cross-validation tests show that mHM is able to capture the streamflow dynamics adequately well across a wide range of climate and physiographical characteristics. The model yields generally better results (with lower spread of model statistics) in basins with higher rain gauge density. Model performance for other fluxes and states is strongly driven by the degree of seasonality that each variable exhibits, with the best match being observed for evapotranspiration, followed by total water storage anomaly, and the least for soil moisture. Results show that constraining the model against streamflow only may be necessary but not sufficient to warrant the model fidelity for other complementary variables. The study emphasizes the need to account for other complementary datasets besides streamflow during parameter estimation to improve model skill with respect to "hidden" variables.

1. Introduction

Since the pioneering work of Crawford and Linsley (1966), the efficiency of computational hydrologic models has been evaluated against streamflow observations that are available at determined locations within a river basin (Dawdy and Lichty 1968; Sorooshian and Dracup 1980; Duan et al. 1992; Bergström 1995; Seibert 2000; Hundecha and Bárdossy 2004; Troy et al. 2008; Yilmaz et al. 2008; Samaniego et al. 2010; Kumar et al. 2013a). This kind of continuous in situ measurement is essential for understanding the governing relationships between rainfall and runoff in a particular drainage basin. The information content of this time series fundamentally differs from other point measurements such as soil moisture and latent heat in the sense that it represents the integral basin response

to a sequence of hydrometeorologic events under particular physiographic and climatic conditions that uniquely characterizes a river basin. Because of this fundamental characteristic, streamflow gauging has been and will be part of the core of national hydrometeorologic monitoring programs and the basis for sound water resources management. It is therefore not surprising that streamflow time series has been the focus for seminal hydrologic work in the past (Kuichling 1889; Sherman 1932; Horton 1935; Hurst 1951; Nash 1958; SCS 1973; Rodriguez-Iturbe and Valdes 1979).

In recent years, however, a tendency toward a more comprehensive assessment of model structural adequacy has taken shape, with an overall aim to improve the representation of different hydrological processes incorporated within a model (Clark et al. 2011; Gupta et al. 2012; Shuttleworth 2012). The rationale behind this assessment is the need to get the right answers for the right reasons (Blöschl 2001; Kirchner 2006), which goes beyond just assessing the model performance against observed streamflow or associated signature measures

Corresponding author address: Luis Samaniego, Helmholtz Centre for Environmental Research–UFZ, Permoserstrasse 15, 04318 Leipzig, Germany.
E-mail: luis.samaniego@ufz.de

(Samaniego and Bárdossy 2007; Yilmaz et al. 2008; Kumar et al. 2010; Pokhrel et al. 2012; Euser et al. 2013). Additional motivation for such assessment is driven by the growing need to simulate spatially distributed land surface fluxes controlled by local soil moisture availability and land surface hydrology. Consequently, complementary datasets representing internal hydrologic states and fluxes, such as soil moisture and evapotranspiration, are required to achieve this goal. New kinds of observations and/or proxy data obtained from remote sensing and/or in situ measurement are becoming increasingly available, although at different spatial and temporal resolutions, for example, monthly total water storage (TWS) anomaly from the Gravity Recovery and Climate Experiment (GRACE) at $1° \times 1°$, near-surface soil moisture from the European Space Agency (ESA) Climate Change Initiative (CCI) at $0.25° \times 0.25°$, and 30-min eddy flux measurements of latent heat with a footprint of hundreds of hectares.

Several recent studies have evaluated the capability of hydrologic and/or land surface models to represent internal model fluxes and/or states (Li et al. 2012; Livneh and Lettenmaier 2012; Sutanudjaja et al. 2014; Cai et al. 2014; Xia et al. 2014, 2015). A common shortcoming in these studies has been the incompatibility of the scales at which simulated state variables and fluxes are compared with the observations (i.e., data are measured at different spatial scales from those at which models usually operate). The scaling issue poses a major obstacle in performing a comprehensive model evaluation (Blöschl 2001; Samaniego et al. 2010; Tetzlaff et al. 2010; Gentine et al. 2012). Often, the in situ measurements of soil moisture or the evapotranspiration inferred at eddy covariance sites are compared with much coarser gridded model outputs (Xia et al. 2014, 2015). In situ measurements often exhibit much finer support than the smallest representative elemental volume of hydrologic models (Blöschl et al. 1995; Wood 1995; Blöschl 1999).

This scale discrepancy problem is exaggerated when a model is evaluated simultaneously against multiple datasets available at different spatial resolutions. In such a case, different up- and downscaling rules have to be employed to enable comparison between simulations and observations. Alternatively, a quasi-scale-independent model parameterization scheme that allows us to reliably represent processes at different spatial resolutions is required to tackle this scaling problem. The latter has the advantage that a process-based modeling approach can be used to estimate hydrologic fluxes and states across multiple scales (Samaniego et al. 2010; Gentine et al. 2012; Kumar et al. 2013b). Most of the existing modeling approaches, however, exhibit scale-dependent performance, which means that the model parameterization obtained at a given spatial resolution induces large bias in hydrologic fluxes and states when applied to other resolutions (Haddeland et al. 2002; Boone et al. 2004; Stöckli et al. 2007; Troy et al. 2008; Samaniego et al. 2010; Kumar et al. 2013b).

Recently, Samaniego et al. (2010) proposed a multiscale parameter regionalization (MPR) method that allows one to make hydrologic predictions at different scales using a same set of model (transfer) parameters but without losing much of the model performance. The method explicitly accounts for the subgrid variability of the essential aspects of the physical processes that are embedded within model parameters (e.g., soil porosity) and ensures that water fluxes simulated at different scales are comparable. The MPR method incorporated within the mesoscale hydrologic model (mHM; Samaniego et al. 2010) has been tested across a variety of climate and land surface conditions at different spatial resolutions ranging from 4 to 100 km (Kumar et al. 2010, 2013a,b; Samaniego et al. 2013). To date, these scaling studies have mainly focused on evaluating model performance against streamflow and conducting flux-matching experiments using modeled variables at multiple scales and locations.

In this study, we specifically evaluate the ability of the MPR method to reproduce the spatiotemporal dynamics of various water fluxes and states observed at multiple resolutions. The model parameterization constrained using streamflow observations across 400 European river basins is evaluated against complementary datasets that include gridded upscaled in situ evapotranspiration (ET) data, satellite-based soil moisture (SM), and TWS anomalies, as well as local-scale eddy covariance data and their native resolutions. Alternative data fusion possibilities (such as data assimilation) to mitigate the limitations of models are beyond the scope of the present study and require future investigation. The multiscale evaluation approach followed here differs from previous hydrological model assessment studies that have covered the European domain using, for example, the LISFLOOD model (Wanders et al. 2014), the PCRaster Global Water Balance (PCR-GLOBWB) model (Wada et al. 2010; Sutanudjaja et al. 2014), or the Water–Global Assessment and Prognosis (WaterGAP) model (Werth and Güntner 2010). Although these studies focused on evaluating model skill on multiple variables, they have been operated on a limited number of basins and/or with little consideration of the scaling discrepancy problem while verifying model outputs against observations.

We hypothesize that parameter estimation based only on streamflow-related metrics is a necessary but not sufficient condition to warrant the proper partitioning of incoming precipitation P into various spatially distributed water storage components (e.g., SM) and fluxes
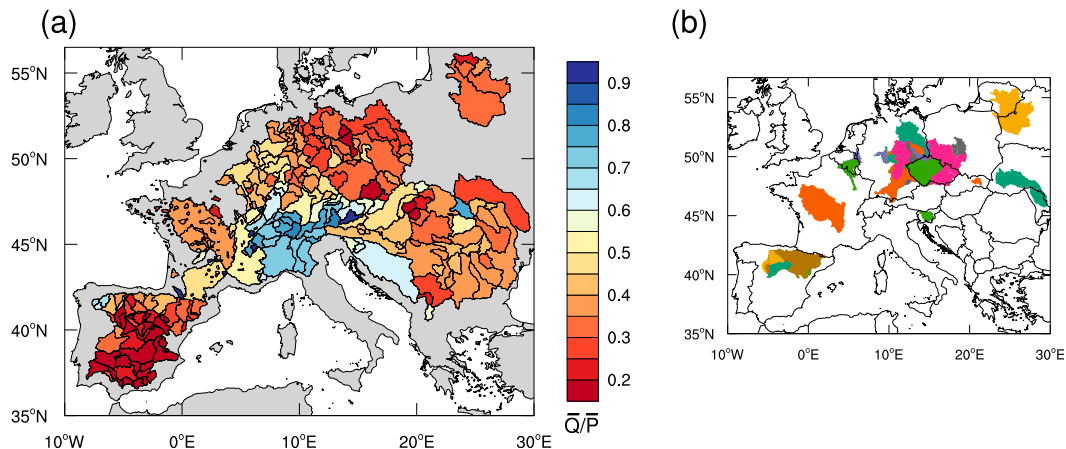
FIG. 1. (a) Spatial map of the modeling domain showing the runoff ratio (i.e., $\overline{Q}/\overline{P}$) for 400 European basins used in this study. The smaller basins are overlaid on larger ones. (b) The 36 donor basins provide an ensemble of plausible parameter sets (i.e., γ) constrained using the observed streamflow (different colors are used to distinguish between individual basins).

(e.g., ET). In the presented study, the multiscale and multivariate verification of water fluxes and states is carried out by executing mHM in a "seamless manner" (i.e., multiscale model simulation, in which each scale realization can be run simultaneously using a same set of model transfer parameters) at the native resolutions of available datasets varying from 0.5 to 100 km.

2. Data and methods

a. Study area and datasets

The study is carried out in 400 European river basins (Fig. 1a) with drainage areas varying from 10^2 to 10^6 km^2. These basins span across distinct climate conditions ranging from the dry summer subtropical (in the Mediterranean, southern Europe) to maritime temperate (in western Europe) and warm summer continental (in eastern Europe) climate types, according to the Köppen–Geiger classification (Rubel and Kottek 2010). Figure 1a shows the span of runoff ratio ($\overline{Q}/\overline{P}$) that represents the long-term average partitioning of the precipitation $\overline{P}$ into runoff $\overline{Q}$ and actual ET ($\overline{ET}$). The runoff ratio is a comprehensive measure of physiographic basin and regional climate descriptors (Berger and Entekhabi 2001; Sankarasubramanian and Vogel 2002) that ranges between 0 and 1. Basins with smaller $\overline{Q}/\overline{P}$ values represent relatively drier conditions with higher evaporative rates (e.g., southern Spain), while larger $\overline{Q}/\overline{P}$ represents humid or mountainous basins with lower evaporation rates (e.g., alpine regions).

The physiographical datasets used to set up the model mainly include digital elevation models, soil textural properties, and land-cover states. An overview of these

datasets is provided in Table 1. Since these datasets are available at different spatial resolutions, they are mapped on a common spatial resolution of 500 m × 500 m. These fine-resolution datasets then allow one to account for subgrid variability of basin physical characteristics in parameter regionalization, as described further in section 2b.

The meteorological forcings for the mHM consist of the daily gridded fields of precipitation and average, maximum, and minimum air temperatures at 0.25° × 0.25° resolution for the period 1950–2010. These datasets are acquired from the European daily high-resolution gridded dataset (E-OBS, version 8.0; Haylock et al. 2008). These fields were created using the external drift kriging interpolation technique from ground-based observation networks. The potential evapotranspiration is derived using the temperature-based method of Hargreaves and Samani (1982) at the same spatial resolution (0.25° × 0.25°).

Streamflow records are commonly used to constrain the model parameterization and to evaluate its performance. Daily streamflow data between 1950 and 2010 were obtained from the Global Runoff Data Centre (GRDC) for this purpose. The data availability varies from station to station, with the median record length of 43 years. All basins used in this study have undertaken a first-order data quality check so that they do not violate the physical constraints imposed by the Budyko relationship (Budyko 1974) and do not exhibit any obvious unnatural behavior in the discharge time series. More detailed analysis, particularly on the degree of regulation of European river basins, is deemed beyond the scope of the study because of the lack of support information. Besides streamflow, model performance is evaluated

TABLE 1. Description of input and evaluation datasets. Acronyms: ECAD, European Climate Assessment and Dataset; SRTM, Shuttle Radar Topography Mission; CGIAR-CSI, Consultative Group on International Agricultural Research–Consortium for Spatial Information; EEA, European Environment Agency; ESD, European Soil Database; and HWSD, Harmonized World Soil Database. Other acronyms are defined in the text.

Variable	Description	Reference
Model setup		
Meteorological forcing inputs (precipitation, air temperature, and potential ET)	Daily E-OBS product at 0.25° × 0.25° resolution	ECAD (http://www.ecad.eu; Haylock et al. 2008)
Terrain characteristics (e.g., elev, slope, aspect, flow direction, and flow accumulation)	SRTM DEM data at 90 m × 90 m resolution	CGIAR-CSI (http://srtm.csi.cgiar.org; Jarvis et al. 2008)
Land cover (e.g., major class: forest, permeable, and impervious cover)	CORINE land-cover dataset at 100 m × 100 m	EEA (http://www.eea.europa.eu)
Soil textural properties (sand and clay content, bulk density, and horizon depth)	30 arc-s raster based on ESD	HWSD (FAO/IIASA/ISRIC/ISSCAS/JRC 2012)
Model evaluation		
Streamflow data	Daily observed streamflow	GRDC (http://www.bafg.de/GRDC; French basins; Giuntoli et al. 2013)
TWS anomaly	Gridded product at 1° × 1° resolution	GRACE (Swenson and Wahr 2006; Landerer and Swenson 2012)
Actual ET	Gridded product at 0.5° × 0.5° resolution In situ observations	FLUXNET (Jung et al. 2011) Eddy covariance sites (http://gaia.agraria.unitus.it)
SM	Gridded product at 0.25° × 0.25° resolution	ESA CCI (http://www.esa-soilmoisture-cci.org; Liu et al. 2011; Dorigo et al. 2014)

against complementary datasets, namely, the TWS anomaly, actual evapotranspiration, and soil moisture. A brief overview of these datasets is given below.

1) TWS ANOMALY

The TWS anomaly represents an important measure of seasonal and interannual variability of the terrestrial water storage and is of critical interest for water resource management. The state of TWS affects infiltration rates, subsurface flows, groundwater recharge, and runoff generation (Li et al. 2012). The remotely sensed anomalies of Earth's gravity field retrieved by GRACE (release 05; Landerer and Swenson 2012) are used in this study to evaluate the simulated TWS of mHM. The global GRACE gridded dataset has 1° × 1° spatial and monthly temporal resolution. Although the GRACE product is available at coarse spatial and temporal resolutions, its application in hydrologic studies is increasing (Andersen et al. 2005; Zaitchik et al. 2008; Su et al. 2010; Li et al. 2012; Forman et al. 2012; Livneh and Lettenmaier 2012; Cai et al. 2014; Orth and Seneviratne 2015). The TWS anomaly is analyzed using a combined product composed of different solutions obtained from three processing centers: the GeoForschungsZentrum (GFZ; Potsdam, Germany), the Center for Space Research at the University of Texas at Austin (United States), and the Jet Propulsion

Laboratory (United States). The TWS anomaly is calculated via removing their corresponding long-term mean estimates, which cover the baseline period from January 2004 to December 2009 (NASA 2015). The arithmetic mean of these three products used here is the most effective way to reduce noise in the gravity field within the available scatter of the three solutions (Sakumura et al. 2014). The evaluation period for the TWS anomaly ranges between 2004 and 2012.

2) ACTUAL ET

Actual ET (latent heat flux) includes evaporation of water from soil, surface water bodies, canopy interception, and transpiration from plants leaves. It represents the second-largest flux of the hydrologic cycle; on average, 60% of terrestrial precipitation is returned back to the atmosphere via ET (Oki and Kanae 2006). In this study, the modeled ET is evaluated against data at two distinct resolutions from 1) finescale eddy covariance observations at 27 CarboEurope sites (Göckede et al. 2008; Mauder et al. 2008) and 2) the 0.5° gridded ET dataset derived from the Flux Network (FLUXNET) observations (Jung et al. 2011).

Basic information for the eddy covariance stations is provided in Table A1 (in the appendix). The footprint of the observations covers approximately 700 m. Only stations with an almost complete record for the years

2004–07 are chosen from the CarboEurope database. Data are processed after Papale et al. (2006), and unit imputation (gap filling) is done by marginal distribution sampling (Reichstein et al. 2005). Observed latent heat fluxes were corrected by authors for missing energy balance closures with a Bowen ratio approach similar to Kessomkiat et al. (2013).

The gridded FLUXNET ET product is acquired from the Department of Biogeochemical Integration at the Max Planck Institute for Biogeochemistry, Jena, Germany. The FLUXNET ET product is obtained by upscaling observations of biosphere–atmosphere fluxes of carbon and energy from eddy covariance flux tower sites using model tree ensembles (MTE; Jung et al. 2011). The global monthly ET product is available at $0.5° \times 0.5°$ for the period 1982–2011. We refer to Jung et al. (2011) for a detailed description of the processing algorithm used to generate this dataset.

3) ESA CCI SURFACE SM

SM acts as a switch and integrator of various energy and water fluxes between the land surface and the atmosphere and is the life-giving substance for vegetation. Correctly estimating the degree of soil saturation is the key point in hydrological modeling because it influences the partitioning of precipitation into ET and runoff. It also has a direct effect on society in terms of agriculture management as well as flood and drought predictions. Moreover, it integrates precipitation and evaporation over periods of days to weeks, thus introducing memory in the hydrological cycle.

The ESA CCI provides a global SM product based on the retrievals from four passive (SMMR, SSM/I, TMI, and AMSR-E) and two active [the ERS Active Microwave Instrument (AMI) and Advanced Scatterometer (ASCAT)] coarse-resolution microwave sensors. The interested reader may refer to Liu et al. (2011) and Dorigo et al. (2014) for detailed descriptions of this dataset. The ESA CCI dataset represents near-surface SM (0.5–2 cm) at $0.25° \times 0.25°$ spatial resolution for the period 1978–2013. The recent study by Dorigo et al. (2014) shows that the skill of the merged product compared to the skill of the individual input products of the passive–active sensors with respect to in situ observation has a comparable and/or better performance than the individual input products in terms of the Spearman rank correlations. We emphasize that the ESA CCI SM product is rescaled to the dynamic range of the GLDAS Noah surface soil moisture fields, and therefore it could not be considered as an independent dataset representing absolute true soil moisture (ESA 2015).

b. The mHM and the MPR

The mHM used in this study is a grid-based distributed model that is grounded on numerical approximations of dominant hydrologic processes applied in known HMs such as the Hydrologiska Byråns Vattenbalansavdelning (HBV; Bergström 1995) model and the Variable Infiltration Capacity model (VIC; Liang et al. 1994). Specifically, the model accounts for the following processes: canopy interception, snow accumulation and melting, SM dynamics, infiltration and surface runoff, ET, subsurface storage and discharge generation, deep percolation and base flow, and flood routing. The snow accumulation and melting processes are modeled using a modified degree-day method that accounts for the enhanced snowmelt during the intense precipitation events (Hundecha and Bárdossy 2004). The incoming precipitation and snowmelt is partitioned into root-zone soil moisture and runoff components, depending on the degree of soil saturation, using a power function similar to the HBV model. The model uses three soil layers to describe the root-zone soil moisture dynamics: the depth of the first soil layer is 5 cm, the second one is 25 cm, and the third layer is up to 100 cm. The soil moisture processes in the first two soil layers account for the variation in soil organic matter over time with changes according to land-cover type. Actual ET from soil layers is estimated as a fraction of potential evapotranspiration depending on the soil moisture stress and fraction of roots in each soil layer. The runoff generation process in mHM accounts for surface, fast- and slow-interflow, and baseflow components. The interflow component represents the fast reaction to weather signals while the base flow represents the slow and permanent groundwater flow. Finally, the total runoff produced at each grid cell is routed to the neighboring downstream grid cell via the Muskingum–Cunge flow routing algorithm. For a complete model description, interested readers may refer to Samaniego et al. (2010; the model code can be downloaded from www.ufz.de/mhm). To date, mHM has been applied over a large number of river basins across Germany, Europe, and the United States (Samaniego et al. 2010, 2013; Kumar et al. 2010, 2013a,b; Livneh et al. 2015; Thober et al. 2015).

The model uses three distinct levels of information to better account for spatial heterogeneities of input data, hydrological processes, and meteorological forcings. The lowest level ℓ_0 describes information on input data related to physiographical and morphological characteristics of a basin. The intermediate level ℓ_1 is used to model the governing hydrological processes, while the highest level ℓ_2 contains information on meteorological datasets. Typically, the spatial resolution of ℓ_1 and ℓ_2 is

the same, on the order of kilometers depending on the availability of the forcing dataset (24 km ≈ 0.25° in this study). The resolution of ℓ_0 input data is much finer than the other two, on the order of hectometers (500 m ≈ 0.004° in this study).

The MPR technique (Samaniego et al. 2010; Kumar et al. 2013b) is used to efficiently incorporate the subgrid-level (ℓ_0) information within the modeling level (ℓ_1) using a two-step parameterization technique (see Fig. 2). In the first step, model parameters β^0 (e.g., porosity) are linked to available basin physical characteristics (e.g., terrain slope or sand and clay contents) using a set of pedotransfer functions f and global parameters γ. This linkage is established at the ℓ_0 spatial resolution to account for the subgrid variability of input data and β^0. In the subsequent step, the ℓ_0 fields of model parameters are aggregated to generate the effective regional parameter fields β^1 at the modeling level (ℓ_1). The aggregation is performed using upscaling operators such as the harmonic, geometric, or arithmetic means, which satisfy flux-matching conditions, that is, minimal discrepancy between aggregated water fluxes simulated across multiple resolutions. A set of global parameters is usually inferred via a suitable parameter estimation technique. This two-step parameter regionalization technique allows the model to run efficiently in a seamless manner at multiple resolutions using the same set of global parameters (see Fig. 2).

According to Gupta et al. (2014, p. 471), the benefit of a regionalization method such as MPR stems from the fact that it "regularize[s] the optimization problem, limiting the degrees of freedom to a small number of regional transfer function coefficients." The MPR technique, in addition to regularizing the optimization problem, takes into account the subgrid variability of the essential aspects of the physical process that represents a given model parameter (e.g., soil porosity, wilting point, or hydraulic conductivity). Previous studies have demonstrated the effectiveness of the MPR approach over other existing parameterization techniques based on hydrological response units, lumped parameterizations, and standard regionalizations that do not account for the subgrid variability of model parameters (Kumar et al. 2010, 2013b; Samaniego et al. 2010, 2011).

c. Experimental design and model setup

The goal here is to comprehensively evaluate the skill of mHM to represent the spatiotemporal variability of modeled fluxes and states at multiple scales. The experimental design of model parameter estimation and seamless verification is schematically shown in Fig. 2. In the initial phase, the model parameters are constrained against observed streamflow to obtain an ensemble of plausible model parameters (i.e., γ) with the following procedure.

1) For each basin, the following steps must be taken:

   (i) Prior to the parameter estimation of the model parameters, a subset of "informative" parameters γ^* is identified using the sequential screening method developed by Cuntz et al. (2015). This screening method is an adaption of the Morris method (Morris 1991). In the first iteration, the model is evaluated at several points along trajectories of the parameter space and elementary effects are determined. Parameters with an elementary effect above a certain threshold are considered to be informative. The next iteration of the method only takes into account "noninformative" parameters to test whether they are sensitive at other regions of the parameter space. This iterative procedure is repeated until no additional parameters are marked to be informative.

   (ii) Find an optimal set of $\hat{\gamma}$ identified in the first step by maximizing the Kling–Gupta efficiency (KGE; Gupta et al. 2009). The shuffled complex evolution (SCE) algorithm (Duan et al. 1993) is used to maximize

   $$\max_{\hat{\gamma}} \mathrm{KGE}(r, \alpha, \beta), \qquad (1)$$

   with

   $$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}, \qquad (2)$$

   where r is the Pearson correlation coefficient between observed Q_{obs} and simulated $Q(\hat{\gamma})$ streamflow; α denotes the measure of relative variability in the simulated and observed values (ratio of the standard deviations); and β is the ratio between the mean simulations and mean observations, that is, bias. The parameter estimation period varies from basin to basin and ranges between 4 and 16 years of data, depending on the availability of the observed streamflow. Prior to the model parameter estimation, a default run in the period from 1951 to 2010 is conducted to ensure appropriate initializations of internal model states and fluxes. Additionally, 5 years of data prior to the parameter estimation period are used to spinup the model. The model is executed at a daily time step and
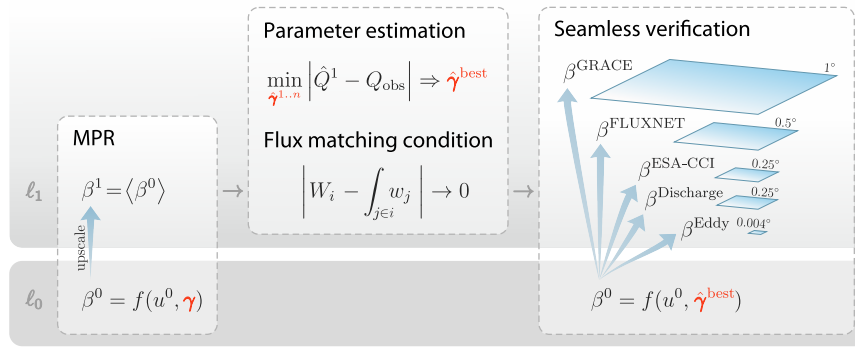
FIG. 2. Schematic of the experimental setup of the multiscale and multivariable model evaluation. (left) The two-step MPR scheme to incorporate the subgrid variability of basin physical characteristics (i.e., u^0) available at ℓ_0. Effective parameters (i.e., β^1) at ℓ_1 are estimated using a set of transfer functions (i.e., f), global (estimation) parameters (i.e., γ), and upscaling operators (i.e., angle brackets $\langle \cdot \rangle$). (middle) The subsequent procedure to estimate a best set of global parameters (i.e., $\hat{\gamma}^{\text{best}}$) calibrated against observed discharge (i.e., Q), and upscaling operators using the water fluxes matching conditions across multiple ℓ_1 resolutions [variables W_i and w_j denote the fluxes estimated at the coarser and finer cells i and j, respectively; for more details, see Samaniego et al. (2010)]. (right) Illustration of the seamless (multiscale) verification approach for water fluxes and states at multiple modeling scales $\ell_1 = 0.004°, \ldots, 1°$, using the same subgrid level information (β^0) and $\hat{\gamma}^{\text{best}}$.

spatial resolution of $0.25° \times 0.25°$. The optimal set of $\hat{\gamma}$ inferred by this procedure corresponds to the complete basin and not to individual grid cells. This step is repeated 10 times with random initialization of the parameter space to partially account for the uncertainty of the SCE algorithm, and the best-performing parameter set is selected.

  (iii) Transfer $\hat{\gamma}$ to all remaining basins for stream-flow cross validation. Estimate KGE($\hat{\gamma}$) as a measure of transfer efficiency.

2) Select the best parameter sets $\hat{\gamma}^{\text{best}}$ from the pool of $\hat{\gamma}$ based on a cross-validation median KGE value larger than a threshold κ.

Following this procedure, an ensemble of the best 36 parameter sets (i.e., $\hat{\gamma}^{\text{best}}$) satisfying the threshold κ of 0.55 is selected to represent the cross-validation uncertainty of model output. The κ criterion of 0.55 is not directly related to performance in an individual basin, but rather, it represents the median KGE value in a cross validation over 400 basins. The location of the respective 36 basins spans over the entire study domain (Fig. 1b), which indicates the representativeness of the donor sample.

In the second phase (Fig. 2, right), the ensemble of 36 parameter sets is used to conduct model simulations at the native scale of the complementary datasets using the following procedure.

For each parameter set $\hat{\gamma}^{\text{best}}$, take the following steps:

1) Estimate effective model parameters (i.e., β^0) at ℓ_0 using transfer functions (i.e., f) applied to the basin's physical characteristics u^0 (Fig. 2, left).
2) Aggregate β^0 to β^1 at multiple modeling levels (i.e., ℓ_1) using upscaling operators (indicated by angle brackets $\langle \cdot \rangle$). The spatial resolution of ℓ_1 varies from $0.004°$ to $1°$ depending on the variable of interest (Fig. 2, right).
3) Run mHM using the β^1 fields at multiple scales and evaluate its performance for selected water fluxes and states.

Following the aforementioned algorithm, the model is executed 36 times at the spatial resolutions of $1° \times 1°$, $0.5° \times 0.5°$, and $0.25° \times 0.25°$ over the entire domain and at $0.004° \times 0.004°$ across 27 eddy grid points (see Fig. 2).

The streamflow evaluation is conducted on the entire record of available streamflow observations within the simulation period 1951–2010. The model skill for complementary gridded datasets is evaluated during the period of 2004–10 for the ESA CCI SM product, 1982–2011 for the gridded LandFlux ET data, and 2004–12 for the GRACE TWS anomaly. The ET evaluation at eddy flux stations is limited to the period 2004–07 because of data availability issues. Three Coordination of Information on the Environment (CORINE) land-cover scenes corresponding to the years 1990, 2000, and 2006 are taken into consideration. Model simulations prior to 1990 use the CORINE 1990 land-cover map.
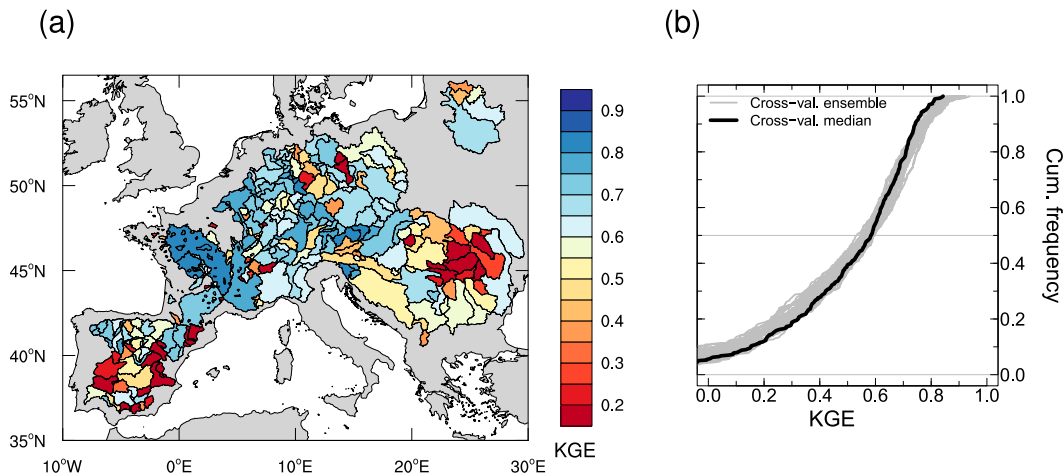
(a)   (b)

FIG. 3. (a) Spatial maps of the modeling domain showing median KGE values between observed and simulated discharge for 400 European basins based on the cross-validation analysis (spatial model resolution of 0.25° × 0.25°; daily time step). (b) Cumulative frequency of the KGE values for the cross-validation uncertainty based on 36 parameter sets (gray), and the median KGE value (black) shown in (a).

The model performance is evaluated using multiple statistical criteria that includes r, α, and β [Eq. (2)], which quantifies a mismatch between model and observations with respect to the temporal dynamics, variability, and biases, respectively. Additionally, the model evaluation results are presented using a Taylor diagram (Taylor 2001). This two-dimensional diagram quantifies concisely how well model simulations match observations in terms of r, α, and the root-mean-square difference.

Finally, a two-dimensional histogram of the marginal distributions of the observed and simulated values (also called empirical copula) is used to describe the statistical dependency between the marginal distributions of two random variables (Nelsen 2006). The copula can be generally written as

$$P(x \leq X, y \leq Y) = C[F(x), G(y)], \qquad (3)$$

where x is the observed quantity with distribution function F, y is the simulated value with distribution function G, the left-hand side of the equation is the joint probability P of x and y, and C is the copula between F and G. The copula-based model evaluation allows us to quantify the stochastic dependence of simulated variables with respect to observations along the entire range of the variable.

3. Results and discussion

a. Model evaluation using observed streamflow

The mHM performance for the median KGE values between observed and simulated Q is depicted in Fig. 3a. The results indicate that mHM is capable of simulating daily discharge well over the pan-European domain considering that around 70% of the total area exhibits a median KGE value exceeding 0.5. It should be noted that the KGE values at a given basin are not obtained by means of onsite parameter estimation, but rather by transferring global parameters (i.e., $\hat{\gamma}$) from other basins. This guarantees consistent representation of hydrological processes at different locations, because the goal here is to obtain parameter sets that represent hydrological fluxes and states across the entire domain applicable for making predictions in ungauged basins. Onsite parameter estimations yield even better performances but increase the dimensionality of the parameter space for the entire domain since each basin in this case requires a set of global parameters. Additionally, Fig. 3b shows the ensemble of the cumulative frequency distributions of model performance based on 36 sets of global parameters (i.e., $\hat{\gamma}^{best}$). This figure illustrates a considerably narrow variation of model performance due to equally well-performing parameter sets with KGE values higher than 0.55 for 50% of the basins.

The model performance in terms of KGE tends to be homogeneously distributed over space (Fig. 3a). Hot spots of poorer model performance occur in notably human-influenced river basins such as those in southern Spain, where the pressure on water resources is high and observed Q is far from natural conditions because of irrigation diversions, hydroelectric power generation, and flood control (Batalla et al. 2004; Lorenzo-Lacruz et al. 2012). The mHM does not include human-influenced processes, as the majority of other rainfall–runoff models;

thus, below-normal performance is expected in those areas. We note that the basins that violate the physical constraints of the Budyko curve are removed prior to the analysis, as discussed earlier; however, this first-order quality check may not be sufficient to filter basins with significant anthropogenic activities. The mHM performance to simulate naturalized streamflow dynamics in other heavily human-influenced U.S. basins is adequate and comparable to other existing models (Kumar et al. 2013a; Livneh et al. 2015). However, the lack of naturalized streamflow dataset in the present study domain limits such types of model evaluation. Another hot spot can be found in eastern Europe (Romanian Carpathian Mountains) where the model systematically underestimates snowmelt-driven floods (in spring). The same behavior is observed by onsite parameter estimation (not shown). In addition to the model conceptual error, the poor performance in these areas can be related to observation errors, such as the precipitation undercatch discussed in the following section.

b. Factors influencing Q predictability

The model performance of KGE and its three components [see r, α, and β in Eq. (2)] is further evaluated in Fig. 4 for basic basin characteristics such as area, rain gauge density, and runoff ratio (i.e., $\overline{Q/P}$). In general, the spread in uncertainty decreases with increasing basin area and the model performance also tends to be improved: KGE and r increase, while α and β converge toward their ideal value of one. This type of model performance dependency indicates that smaller basins are more susceptible to errors in model inputs than the larger ones, which also stems from averaging and the central limit theorem. The closer to the representative elementary area (REA), the more difficult it is to model because of the increased effect of small processes considered neither in the model nor in the data. Such kinds of model dependency are also reported in previous studies (Reed et al. 2004; Merz et al. 2009; Kumar et al. 2013b).

Additionally, Fig. 4 depicts the relation between model performance and rain gauge density (number of rain gauges per $1000 \, km^2$) to investigate the effect of forcing uncertainty. The model exhibits systematically better performance in regions with relatively higher rain gauge density, particularly in terms of variability. This promotes the importance of having a dense observation network for meaningful hydrological simulations, which is in particular important for capturing small-scale features such as convective cells (Alfieri et al. 2014). The median rain gauge density is 0.4 gauges per $1000 \, km^2$ and does not meet the standard of the World Meteorological Organization (WMO) in which the tolerable rain gauge

density in flat regions is around 1–2 gauges per $1000 \, km^2$, while it increases to 4–10 gauges per $1000 \, km^2$ for mountainous regions (Dingman 2004). Therefore, a precipitation product based on sparsely distributed rain gauge data can lead to higher modeling errors arising from imperfect precipitation estimates. However, we would like to emphasize that E-OBS used here is the best possible freely available dataset that exists at the moment with a relatively long temporal coverage, large spatial extent, and fine spatial resolution (Hofstra et al. 2009).

Moreover, Fig. 4 illustrates that the model performance with respect to $\overline{Q/P}$ is usually superior in intermediate physiographic and climatic regimes ($0.3 < \overline{Q/P} < 0.7$), whereas the performance deteriorates toward both extremes. Generally, the model tends to overestimate the observed mean and variability in relatively moisture-limited (dry) basins. Note that these basins contain areas with human-influenced activities where the model performs poorly (as discussed before). On the other hand, the extreme energy-limited basins exhibit a large bias and a systematically underestimated variance. These shortcomings can be related to several factors: precipitation underestimation due to lower rain gauge density, insufficient evaporation rates, and/or model deficiency in capturing subgrid snow processes. In general, the spread of model statistics (r, α, and β) is considerably lower and the model yields better results in basins with higher rain gauge density, which is observed regardless of the selected basin characteristics (cf. gray filled circles with black crosses in Fig. 4 and their median values).

c. Spatial evaluation using complementary data

The model is further evaluated against the following complementary data (not being used to constrain the model) at monthly temporal resolution and native spatial resolution, namely, fields of the TWS anomaly from GRACE ($1° \times 1°$), ET from FLUXNET ($0.5° \times 0.5°$), and SM from ESA CCI ($0.25° \times 0.25°$). Figure 5 shows the model performance in terms of median r of the original data (Fig. 5, top), median r of standardized anomalies (Fig. 5, middle), and corresponding copulas of the latter one (Fig. 5, bottom).

Overall, the model represents the TWS anomaly and the actual ET adequately well, while the performance for SM is not satisfactory in terms of correlation for the original time series (Figs. 5a–c). Presented hydrological variables exhibit strong seasonality and the performance criteria based on, for example, correlation coefficient for such variables is not adequate to show the actual model skill. Therefore, standardized anomalies
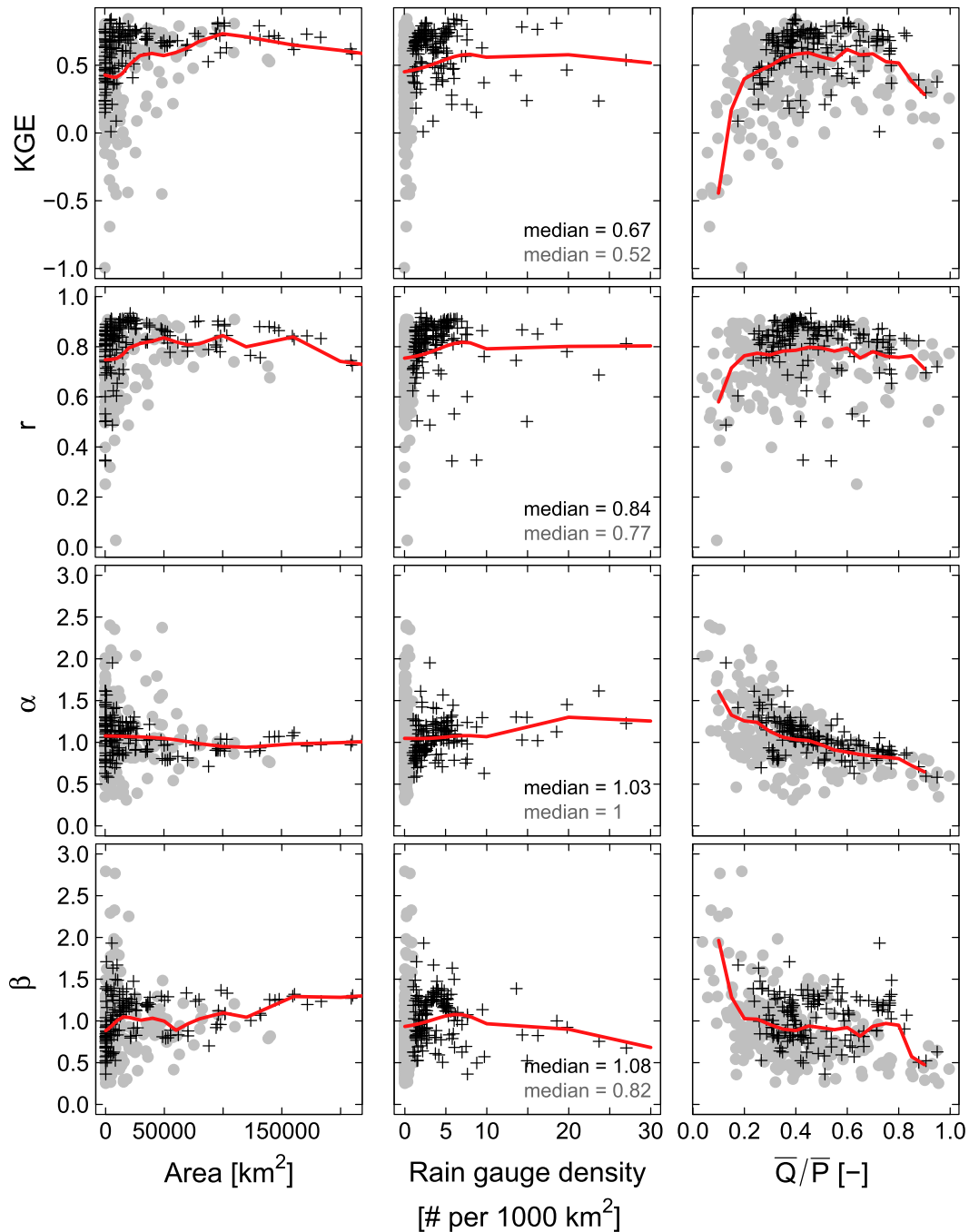
FIG. 4. Basin area, rain gauge density, and runoff ratio as factors influencing model predictability of discharge in terms of KGE (see also Fig. 3a) and decompositions into the three components [r, α, and β in Eq. (2)]. Black plus signs show basins that have at least one rain gauge per 1000 km^2, gray filled circles show otherwise, and median values of the y axis are provided in corresponding colors. Kernel regression is used to produce a smooth red line for the whole sample (black and gray).

of both observations and simulations are estimated by removing their respective monthly means and standard deviations. The correlation for the standardized anomalies shows, in general, deterioration for TWS and ET, but slight improvement in SM when compared to the data with retained seasonality (Figs. 5d–f). The mHM results are consistent in findings of the recent study by Orth and Seneviratne (2015).
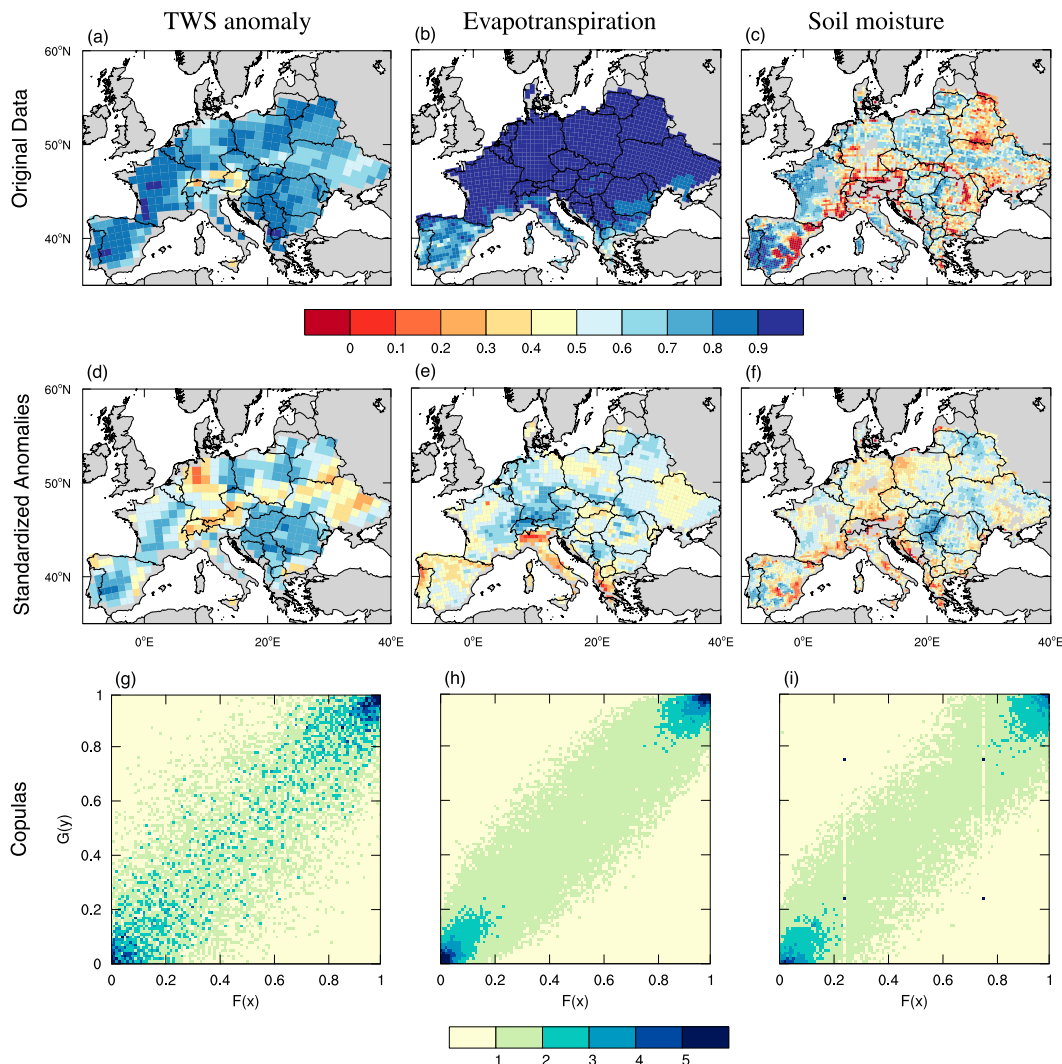
FIG. 5. Model performance of mHM simulations and (left) TWS anomaly (GRACE), (middle) actual ET (gridded FLUXNET), and (right) SM (ESA CCI) observations in terms of medians of the Pearson correlation coefficient of the (a)–(c) original time series, (d)–(f) standardized anomalies, and (g)–(i) empirical copula densities of the standardized anomalies. Term $F(x)$ is a distribution function of an observed variable and $G(y)$ is a distribution function of a simulated variable [see Eq. (3)]. The Spearman rank correlations are 0.61 for TWS anomaly, 0.55 for ET, and 0.49 for SM.

A reasonably good agreement is achieved for TWS anomaly in a large part of the study domain. A relatively larger error can be observed in the Alps and coastal areas. This can be attributed to the fact that mHM lacks the capability to explicitly represent the glacial and tidal processes in these areas. Also, the GRACE data are not suitable to accurately quantify ice mass changes in glaciers (Jacob et al. 2012) and there are relatively higher measurement and leakage errors provided in the GRACE dataset along the coastal line. The leakage error of GRACE stands for the "residual errors after filtering and rescaling" from the raw original product to estimate the TWS anomaly (NASA 2015).

The temporal dynamics of modeled ET resembles quite well the FLUXNET-derived ET product with the majority (75%) of cells exceeding correlation coefficients (i.e., r) larger than 0.9 for the original time series, and larger than 0.59 for the standardized anomalies. Poor model performance is noticed on the Iberian Peninsula. Notably, in these areas the model performance for observed Q is also poor (Fig. 3).

In comparison with the two aforementioned variables, r for soil moisture exhibits the poorest performance; however, overall correlation is quite comparable to those obtained in other recent studies (Dorigo et al. 2014; Lievens et al. 2015). Stripes in the model performance

may correspond to an artifact of the retrieval algorithm, which is a typical characteristic in the observation through satellite microwave instruments (ESA 2015). Notably, the comparison between satellite-derived soil moisture and distributed models is challenging because of data and modeling reasons. First, there is limited information on the exact depth of the soil layer that is used for the ESA CCI satellite product besides other potential retrieval problems such as vegetation coverage, snow, and ice content. Second, the biases across the statistical moments are very typical for surface soil moisture data derived from satellite retrievals, ground measurements, and models so that they need to be quantified and corrected (Reichle and Koster 2004). Third, the top thin first soil layer of most distributed models is not a good representation of actual soil moisture. Water evaporates as vapor from the soil surface. Soil water models that calculate only liquid water flow have to compensate for the missing process of vapor transport. Most of the models therefore include a very thin upper layer that counteracts the liquid water flow. Soil moisture products such as from cosmic-ray sensors (Zreda et al. 2012) represent larger soil volumes (Köhli et al. 2015) and could become more feasible for model–data comparison in future studies.

The overall model evaluation using empirical copula densities of the standardized anomalies shows a strong statistical dependency between observations and simulations in particular for high and low quantiles (Figs. 5g–i). This has strong implications for drought and flood monitoring using satellite products alone. The statistical dependency for values in between the extremes is close to the diagonal but with considerable spread for the three variables analyzed. This is related to grid cells exhibiting low correlation coefficients between their simulated and observed anomalies (Figs. 5d–f). Among the three variables, the relationship between the observed and simulated TWS anomaly exhibits larger scatter because of the reduced sample size due to a coarser spatial resolution and limited temporal availability of datasets (Fig. 5g). Overall, the copula densities indicate that the matching between observations and simulations needs to be improved for normal conditions as compared to the extremes. The Spearman rank correlations estimated based on these copulas vary from 0.49 to 0.61.

d. Basin-scale evaluation of modeled fluxes and states

The quantitative evaluation of model performance at basin scale is presented with Taylor diagrams (Fig. 6) for monthly estimates of Q, TWS anomaly, ET, and SM. The basins are further classified into three categories based on runoff coefficients (i.e., $\overline{Q}/\overline{P}$) representing wet and/or mountainous, intermediate, and dry climatic

regimes. In general, the model is able to represent the temporal dynamics of observed Q adequately well with correlations varying between 0.75 and 0.95 in the majority of the analyzed basins. The observed variability is also well captured by the model regardless of the $\overline{Q}/\overline{P}$ characteristics with a median α value of around 1. On the other hand, the variability in ET is systematically underestimated ($\alpha < 1$), while temporal dynamics are well represented with r exceeding 0.8 in the majority of basins. Furthermore, the performance of the model is relatively low for the total water storage anomaly and SM. This is observed independently of the runoff ratio. The correlation for the TWS anomaly ranges mostly between 0.6 and 0.9 with higher values being noticed for basins lying in the water-limited regime, which is also seen in Fig. 5. The poorest performance among all analyzed variables is observed for simulating the soil moisture dynamics with r less than 0.6 in the majority of basins.

The model performance in terms of correlation for the original time series (shown in Fig. 6) is contrasted against their corresponding standardized values for different variables in Fig. 7a. All variables exhibit lower performance for their standardized estimates with the exception of soil moisture. The largest deterioration is noticed for the ET followed by the TWS anomaly and the least for Q. The sequence of this deterioration corresponds to the degree of seasonality among the analyzed variables. The best model performance for Q can be partly attributed to the fact that the model is constrained against this variable. Despite the lower model performance for standardized variables, the majority of basins have r values above 0.4, which is well beyond the threshold limit of 0.2 to be statistically significant at the 95% confidence interval.

Based on the results shown in Fig. 7a, the test of differences in mean skill scores between the standardized distribution of Q and other modeled variables have p values lower than 10^{-5}. This indicates that the null hypothesis that Q alone can sufficiently constrain model components responsible for internal fluxes and states in a cross-validation mode can be safely rejected. To further support the aforementioned hypothesis, Fig. 7b differentiates the results of the standardized anomalies into two equally sized groups based on the median KGE cutoff value of 0.55 (as discussed in section 3a). On average, a significant deterioration in model skill score (p value <0.01) is observed for other complementary variables in comparison to discharge, with the exception of TWS anomaly, for which no conclusive deterioration in model skill can be noticed for the worse-performing basins. The deterioration is more pronounced for the group of basins yielding on average better model
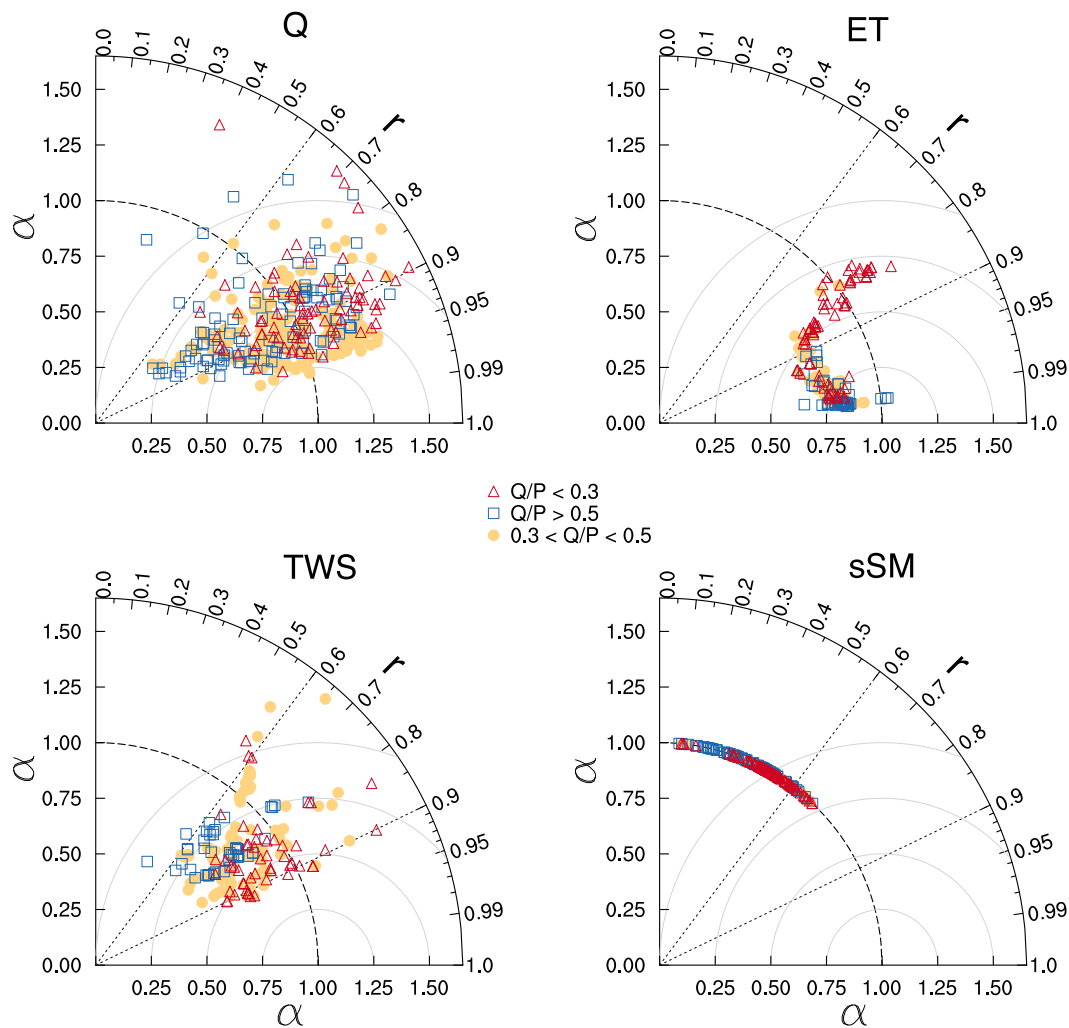
FIG. 6. Evaluating model performance in terms of *r* and *α* for *Q*, actual ET, TWS anomaly, and standardized SM observations at monthly time step using Taylor diagrams. Data are normalized by the std dev of the observations and classified according to the runoff ratio coefficient (i.e., $\overline{Q/P}$).

performance in terms of discharge, which to some degree reflects the overfitting of model parameters during the parameter estimation against the observed discharge. This also indicates that other complementary data are required to appropriately constrain the model in well-performing basins. Such datasets are also of great usage in data-scarce regions, where streamflow observations are not available to constrain the model.

Finally, Fig. 8 shows the monthly dynamics of observed and simulated fluxes and states for three randomly selected basins in dry, intermediate, and wet climatic conditions based on the runoff coefficients. The magnitude and timing of *Q* are well matched and observations are mostly covered within the uncertainty bounds. Contrary to *Q*, ET and TWS anomaly exhibit very regular interannual variability. The model is able to follow this

behavior quite well, although it tends to underestimate the gridded FLUXNET ET in the intermediate and wet basins. The largest discrepancy between the model and observation occurs for SM consistent with our previously discussed results.

e. Model errors in relation to water balance closure

The aforementioned results of basin-scale model performance for different variables illustrate the existence of potential errors between observations and simulations (Figs. 5–8). They can be attributed to a number of factors, which can be mainly related to model and input data errors. The former constitutes errors due to the improper model structure and/or parameterizations, whereas the latter represents errors due to imperfect forcings and/or response variables. An analysis is
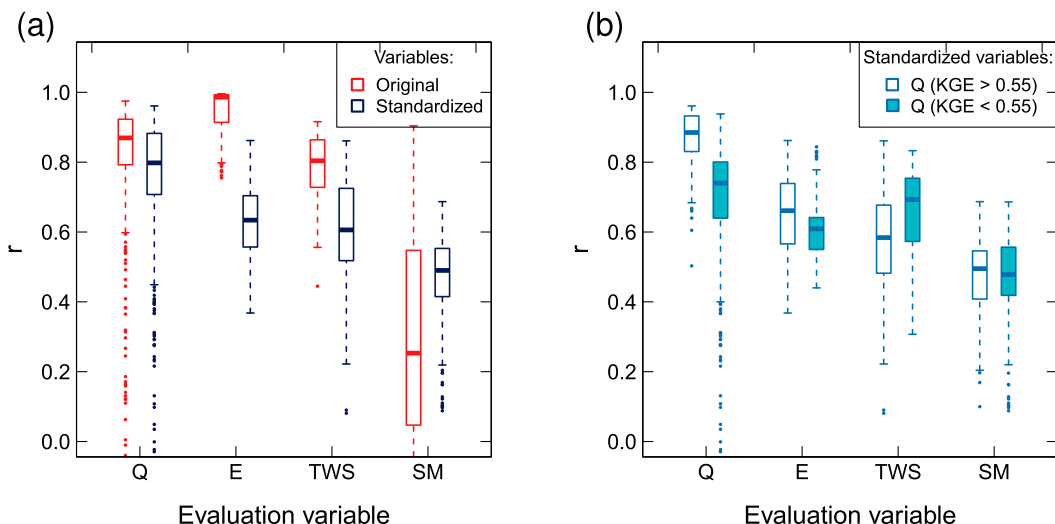
FIG. 7. Evaluating model performance for Q, actual ET, TWS anomaly, and SM at monthly time step using the correlation coefficient. (a) Correlation is derived for the original time series (red; identical to values shown in Fig. 6), and the time series with removed annual cycle normalized by long-term std dev from the monthly data (blue). (b) Results of the standardized anomalies are differentiated into two equal-sized groups based on the KGE values yielding better-performing basins (above median KGE) and worse-performing basins (below median KGE).

carried out to understand the relationship between the errors in input data and errors in individual variables. The residuals in water balance closure $(\overline{P} - \overline{Q} - \overline{ET})$ based on 20 water years (1989–2008) of observed P, Q, and the gridded FLUXNET ET, are taken as a proxy for the input data error. This error is analyzed across basins with different climatic characteristics and contrasted against errors in simulated Q and ET.

Overall, the errors in water balance closure are rather independent from the physiographical characteristics
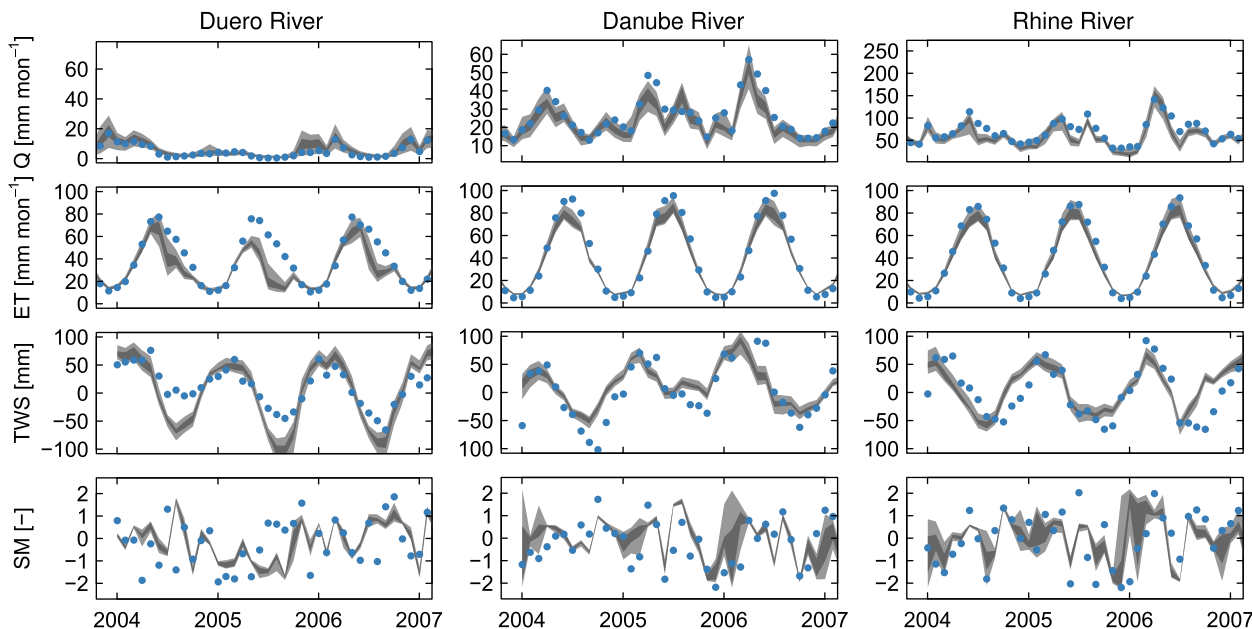

FIG. 8. Time series of monthly Q, actual ET, TWS anomaly, and SM standardized anomaly for three randomly selected basins: (left) the Duero River at Toro (basin area $\approx 42\,000\ \mathrm{km}^2$; $\overline{Q}/\overline{P} = 0.18$), (middle) the Danube River at Zimnicea (basin area $\approx 660\,000\ \mathrm{km}^2$; $\overline{Q}/\overline{P} = 0.41$), and (right) the Rhine River at Basel–Rheinhalle (basin area $\approx 36\,000\ \mathrm{km}^2$; $\overline{Q}/\overline{P} = 0.72$). Observations are shown in blue and mHM simulations are shown given the cross-validation uncertainty with its 95% confidence bounds (light gray) and interquartile range (dark gray). Note the different scales for Q.
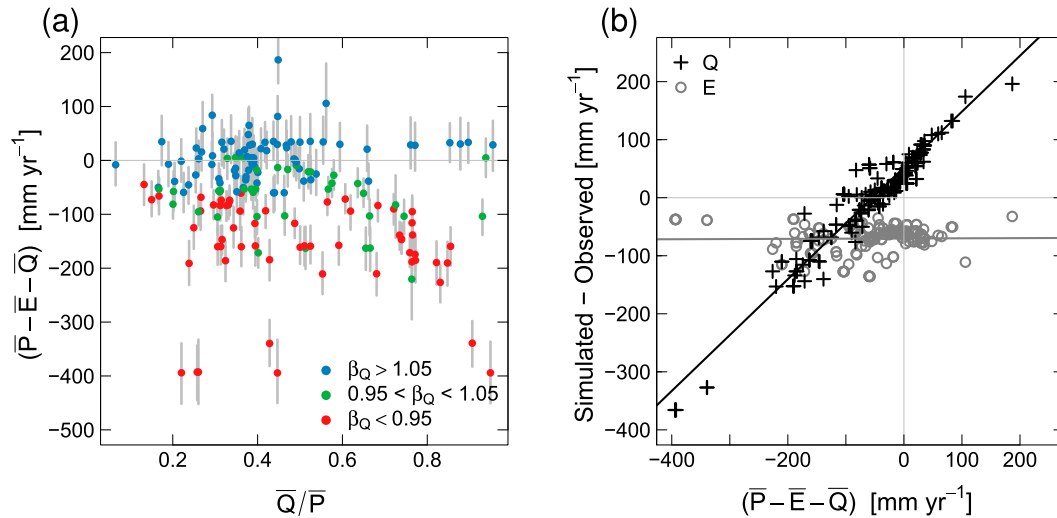
FIG. 9. Analysis of water balance closure error for the 179 basins with full coverage of observed data for the water years of 1989–2008: (a) scatterplot between the runoff ratio and median annual water balance closure error with 95% confidence bounds estimated using the bootstrapping method using 1000 bootstrap samples (gray lines). Color indicates discharge model performance in terms of bias. (b) Relation between median annual water balance closure error and bias between model and observation for streamflow (black) and evapotranspiration (gray).

with a majority of the basins having average annual values between -200 and $100 \, \mathrm{mm} \, \mathrm{yr}^{-1}$ (Fig. 9a). The negative water balance errors are caused by either an underestimated source term (P) or an overestimated sink term (ET + Q). Note that ET is not directly measured but is rather estimated by upscaling observations of biosphere–atmosphere fluxes of carbon and energy from eddy covariance flux tower sites with its own uncertainties (Jung et al. 2011). As noted by Velpuri et al. (2013), the errors in ET can yield up to 50% of the mean annual ET values in certain regions. The underestimation of $\overline{P}$ is likely due to inadequate representation of rain gauge coverage failing to capture the small-scale convective events.

The model overestimates the observed discharge in basins where the positive water balance closure errors occur ($\beta_Q > 1.05$; Fig. 9a), while underestimations are observed in basins with negative closure errors ($\beta_Q < 0.95$; Fig. 9a). This is further supported in Fig. 9b, which indicates that the water balance closure error follows a close relation with the errors between observed and simulated discharge, with a correlation coefficient of around 0.96. The ET errors do not exhibit any dependency to observed water balance errors, since the correlation coefficient is 0.02. The simulated ET estimates averaged across the investigated basins are consistently underestimated by approximately $70 \, \mathrm{mm} \, \mathrm{yr}^{-1}$ with respect to observations. Furthermore, the slope of the best-fitted line between water balance closure error and Q error is nearly one, meaning that, on average, a

100-mm water balance closure error would translate to around 100 mm of simulated streamflow error. The slope can also be interpreted as the elasticity of the fitted line here illustrating the sensitivity of proportional changes in modeling error (in case Q) to the changes in water balance closure error. Results of this analysis indicate that a substantial part of the error in modeled variables can be safely attributed to the erroneous observational datasets. These results highlight the need for better quantification of model errors together with erroneous observational data.

f. Local-scale evaluation of ET

The multiscale evaluation of mHM is further carried out against daily ET estimated at eddy covariance stations with distinct vegetation cover. Results of this analysis, summarized in Fig. 10a, indicate that the model is able to capture the temporal dynamics of ET with correlations ranging between 0.6 and 0.9 across 27 eddy covariance stations. In analogy to the gridded-scale ET simulation results (see Fig. 6), the model systematically underestimates the observed variability, indicating the lack of the model to represent the observed range of ET dynamics (i.e., from dry to wet phases). Figures 10b–10d show the time series of observed and modeled ET at three distinct locations. The temporal dynamics of observations is well represented by the model at the forest and grassland sites with correlations of more than 0.88. Relatively poor performance is observed at the semiarid savanna site ($r = 0.74$). However, the model is able to
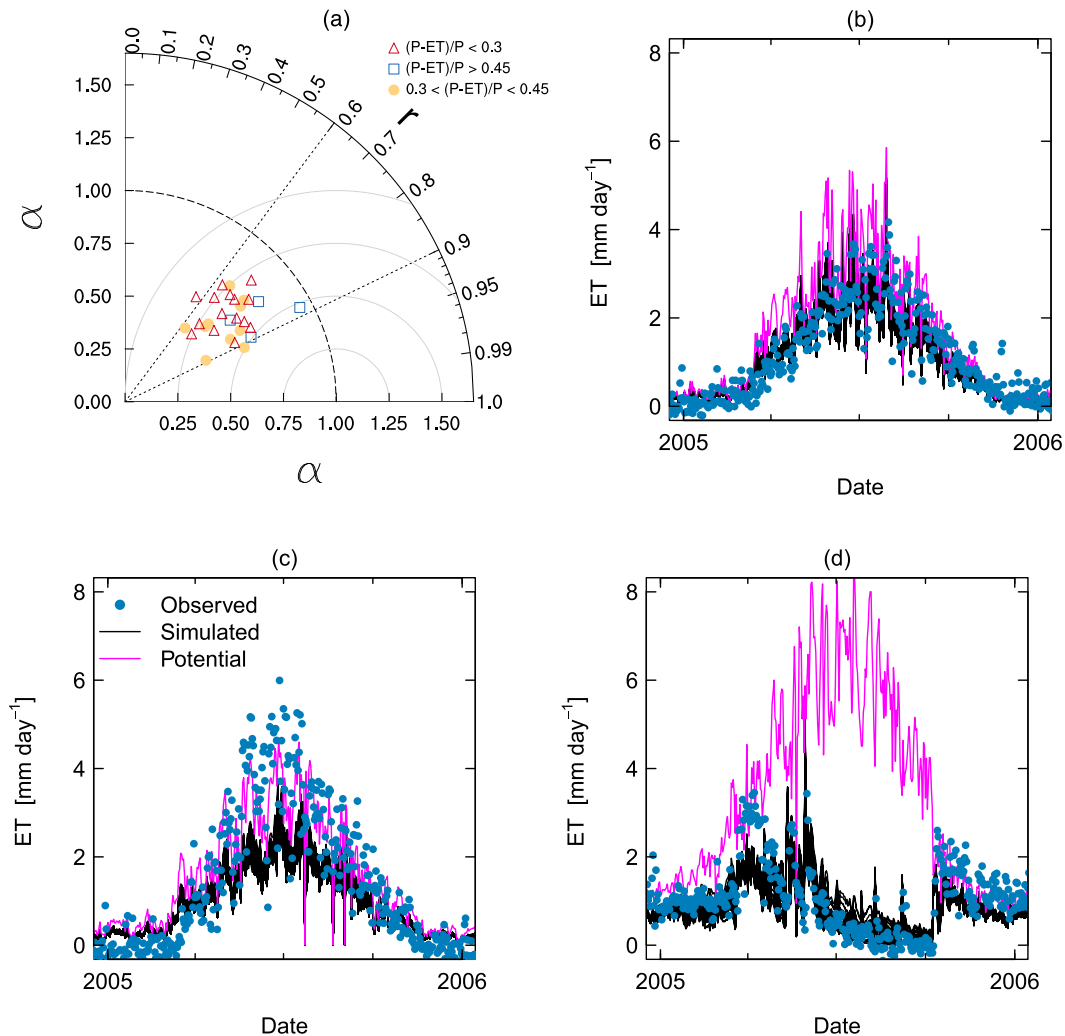
FIG. 10. (a) Comparison of the daily actual ET between mHM and local-scale estimated ET derived at 27 eddy stations (see Table A1 in the appendix) between 2004 and 2007 using Taylor diagrams. Three daily ET time series are shown for (b) evergreen needleleaf forest (Tharandt–Anchor in Germany), (c) grassland (Monte Bondone in Italy), and (d) savanna (Mitra II in Portugal). The mHM simulations are shown for 36 best parameter sets (black).

capture the observed variability, including the sudden jumps in ET values observed at this savanna site. Furthermore, the observed magnitude and variability of ET at the grassland site is strongly underestimated by the model, particularly during summer. This is because the potential ET that is used to force the model at a local scale is generally lower than the observed actual ET.

Results of this analysis provide a first-order confidence that the parameter estimates obtained at much coarser scales can be transferred to finer ones. There is, however, a number of factors that influence the modeling results at the local scale, mainly related to the representation of hydrological processes as well as input data. For example, constraining the model against the local forcings instead of large-scale E-OBS meteorological forcings may improve its performance. Another

limiting factor could be due to the estimation of temperature-based potential ET estimates (Hargreaves and Samani 1982), which do not account for other environmental factors, such as wind speed or humidity (Cristea et al. 2012). Finally, the current model version does not account for lateral flows, particularly relevant at the small scales.

4. Conclusions

The performance of the mesoscale hydrologic model (mHM) parameterized with the multiscale parameter regionalization (MPR) technique is comprehensively evaluated against various in situ and satellite-based observations over 400 European river basins. The multiscale evaluation of internal model fluxes and states is

carried out at the native resolution of available data varying from 0.5 to 100 km using an ensemble of cross-validated model parameters constrained only against observed streamflow. Results show that the model is able to perform well for simulating daily discharge over a wide range of climatic and physiographic conditions with KGE greater than 0.55 in more than 50% of the basins. The streamflow predictability deteriorates in basins with a poor rainfall gauge network and in heavily regulated river basins (e.g., southern Spain). Besides the improvement needed in observational networks, further efforts are needed to incorporate large-scale reservoir operations, irrigation, and other human-induced water withdrawal and storage activities.

The multiscale evaluation for the complementary datasets generally shows reasonable but lower performance in comparison to streamflow, which is used to constrain the model parameters. The model shows the best agreement with the gridded FLUXNET evapotranspiration ($r > 0.8$), followed by the GRACE total water storage anomaly ($0.6 < r < 0.9$) and the least for the ESA CCI merged soil moisture ($r < 0.6$). This performance is strongly related to the degree of seasonality that the selected variable exhibits. The skill of the model deteriorates when the annual cycle is removed from each variable except for the soil moisture, with a majority of the basins exhibiting $r > 0.4$ for the deseasonalized complementary datasets. The analysis of water balance closure errors indicates that a part of the error in modeled variables is due to erroneous observational datasets. While the error between the observed and simulated discharge is closely related to the errors in the water balance closure estimates, modeled ET is consistently underestimated with respect to observations on average by a constant error of $70 \, \text{mm} \, \text{yr}^{-1}$.

The local-scale evaluation of evapotranspiration at several eddy covariance sites further supports the functionality of multiscale parameterization of mHM. While the model is able to capture the temporal dynamics of observed evapotranspiration at most of the sites, it consistently underestimates the observed variability regardless of the locations. Besides improvement in the model parameterization to account for the local-scale processes in detail (e.g., subgrid variability of snow and runoff generation processes), future studies may focus on further enhancement in model performance by constraining the model with site-specific information.

This study provides first-order confidence on the ability of the mHM to simulate fluxes and states across a range of spatial scales and varying climatic and physiographic conditions. Because of the implemented MPR technique, it has been possible to run the model

at disparate scales native to the observational data, without recalibrating the model. Although the model yields good performance while conditioned on observed discharge, further improvements are expected by optimally exploiting other reliable complementary datasets together with the streamflow. Results of this study indicate that the null hypothesis, that streamflow alone can sufficiently constrain model components responsible for internal fluxes and states in a cross-validation mode, can be safely rejected. Therefore, further research should focus on multivariate parameter estimation or assimilation schemes for improving the ability to predict the regional water fluxes and states over large domains.

APPENDIX

Details on the Eddy Covariance Stations

Table A1 provides detailed information on the eddy covariance stations.

TABLE A1. Overview of the eddy covariance stations including land cover, geographic coordinates, and periods used in this study.

No.	Site name (code)[a]	Land cover[b]	Lat	Lon	Available period
1	Neustift (AT-Neu)	GRA	47.116 67°N	11.3175°E	2004–07
2	Brasschaat (BE-Bra)	MF	51.3092°N	4.520 56°E	2004–07
3	Lonzee (BE-Lon)	CRO	50.5522°N	4.744 94°E	2004–07
4	Vielsalm (BE-Vie)	MF	50.3055°N	5.996 83°E	2004–07
5	Oensingen (CH-Oe1)	GRA	47.2856°N	7.732 14°E	2004–07
6	Oensingen (CH-Oe2)	CRO	47.2863°N	7.734 33°E	2004–07
7	Grillenburg (DE-Gri)	GRA	50.9495°N	13.5125°E	2004–07
8	Hainich (DE-Hai)	DBF	51.0793°N	10.452°E	2004–07
9	Mehrstedt (DE-Meh)	GRA	51.2753°N	10.6555°E	2004–06
10	Tharandt (DE-Tha)	ENF	50.9636°N	13.5669°E	2004–07
11	Wetzstein (DE-Wet)	ENF	50.4535°N	11.4575°E	2004–07
12	Soroe (DK-Sor)	DBF	55.4869°N	11.6458°E	2004–07
13	Las Majadas (ES-LMa)	SAV	39.9415°N	5.7734°W	2004–07
14	Vall d'Alinya (ES-VDA)	GRA	42.1522°N	1.4485°E	2004–07
15	Grignon (FR-Gri)	CRO	48.844°N	1.952 43°E	2004–07
16	Hesse (FR-Hes)	DBF	48.6742°N	7.064 62°E	2004–07
17	Le Bray (FR-LBr)	ENF	44.7171°N	0.7693°W	2004–07
18	Bugac (HU-Bug)	GRA	46.6911°N	19.6013°E	2004–07
19	Matra (HU-Mat)	GRA	47.8469°N	19.726°E	2004–06
20	Lavarone (IT-Lav)	ENF	45.9553°N	11.2812°E	2004–06
21	La Mandria (IT-LMa)	GRA	45.5813°N	7.154 63°E	2004–06
22	Monte Bondone (IT-MBo)	GRA	46.0156°N	11.0467°E	2004–07
23	Renon (IT-Ren)	ENF	46.5878°N	11.4347°E	2004–07
24	Roccarespampani 2 (IT-Ro2)	DBF	42.3903°N	11.9209°E	2004–07
25	San Rossore (IT-SRo)	ENF	43.7279°N	10.2844°E	2004–07
26	Loobos (NL-Loo)	ENF	52.1679°N	5.743 96°E	2004–07
27	Mitra IV (PT-Mi2)	GRA	38.4765°N	8.0246°W	2004–07

[a] First two capital letters of the code stand for the country code.

[b] Abbreviations: GRA, grassland; MF, mixed forest; CRO, cropland; DBF, deciduous broadleaf forest; ENF, evergreen needles forest; and SAV, savanna.

REFERENCES

Alfieri, L., F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, and P. Salamon, 2014: Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.*, **517**, 913–922, doi:10.1016/j.jhydrol.2014.06.035.

Andersen, O. B., S. I. Seneviratne, J. Hinderer, and P. Viterbo, 2005: GRACE-derived terrestrial water storage depletion associated with the 2003 European heat wave. *Geophys. Res. Lett.*, **32**, L18405, doi:10.1029/2005GL023574.

Batalla, R. J., C. M. Gómez, and G. Kondolf, 2004: Reservoir-induced hydrological changes in the Ebro River basin (NE Spain). *J. Hydrol.*, **290**, 117–136, doi:10.1016/j.jhydrol.2003.12.002.

Berger, K. P., and D. Entekhabi, 2001: Basin hydrologic response relations to distributed physiographic descriptors and climate. *J. Hydrol.*, **247**, 169–182, doi:10.1016/S0022-1694(01)00383-3.

Bergström, S., 1995: The HBV model. *Computer Models of Watershed Hydrology*, V. P. Singh, Ed., Water Resources Publications, 443–476.

Blöschl, G., 1999: Scaling issues in snow hydrology. *Hydrol. Processes*, **13**, 2149–2175, doi:10.1002/(SICI)1099-1085(199910)13:14/15<2149::AID-HYP847>3.0.CO;2-8.

——, 2001: Scaling in hydrology. *Hydrol. Processes*, **15**, 709–711, doi:10.1002/hyp.432.

——, R. B. Grayson, and M. Sivapalan, 1995: On the representative elementary area (REA) concept and its utility for distributed rainfall–runoff modelling. *Hydrol. Processes*, **9**, 313–330, doi:10.1002/hyp.3360090307.

Boone, A., and Coauthors, 2004: The Rhône-Aggregation land surface scheme intercomparison project: An overview. *J. Climate*, **17**, 187–208, doi:10.1175/1520-0442(2004)017<0187:TRLSSI>2.0.CO;2.

Budyko, M., 1974: *Climate and Life*. D. H. Miller, Ed., International Geophysics Series, Vol. 18, Academic Press, 508 pp.

Cai, X., Z.-L. Yang, C. David, G.-Y. Niu, and M. Rodell, 2014: Hydrological evaluation of the Noah-MP land surface model for the Mississippi River basin. *J. Geophys. Res. Atmos.*, **119**, 23–38, doi:10.1002/2013JD020792.

Clark, M. P., D. Kavetski, and F. Fenicia, 2011: Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.*, **47**, W09301, doi:10.1029/2010WR009827.

Crawford, N. H., and R. K. Linsley, 1966: Digital simulation in hydrology: Stanford Watershed Model IV. Tech. Rep. 39, Dept. of Civil Engineering, Stanford University, 210 pp.

Cristea, N. C., S. K. Kampf, and S. J. Burges, 2012: Revised coefficients for Priestley–Taylor and Makkink–Hansen equations for estimating daily reference evapotranspiration. *J. Hydraul. Eng.*, **18**, 1289–1300, doi:10.1061/(ASCE)HE.1943-5584.0000679.

Cuntz, M., and Coauthors, 2015: Computationally inexpensive identification of noninformative model parameters by sequential screening. *Water Resour. Res.*, **51**, 6417–6441, doi:10.1002/2015WR016907.

Dawdy, D. R., and R. W. Lichty, 1968: Methodology of hydrologic model building. *IAHS Publ.*, **81**, 347–355. [Available online at http://iahs.info/uploads/dms/iahs_081_0347.pdf.]

Dingman, S., 2004: *Physical Hydrology.* Prentice-Hall, 646 pp.

Dorigo, W., and Coauthors, 2014: Evaluation of the ESA CCI soil moisture product using ground-based observations. *Remote Sens. Environ.*, **162**, 380–395, doi:10.1016/j.rse.2014.07.023.

Duan, Q., S. Sorooshian, and V. Gupta, 1992: Effective and efficient global optimization for conceptual rainfall–runoff models. *Water Resour. Res.*, **28**, 1015–1031, doi:10.1029/91WR02985.

——, V. Gupta, and S. Sorooshian, 1993: Shuffled complex evolution approach for effective and efficient global minimization. *J. Optim. Theory Appl.*, **76**, 501–521, doi:10.1007/BF00939380.

ESA, 2015: Soil moisture CCI. European Space Agency, accessed 1 July 2015. [Available online at http://www.esa-soilmoisture-cci.org/node/136.]

Euser, T., H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G. Savenije, 2013: A framework to assess the realism of model structures using hydrological signatures. *Hydrol. Earth Syst. Sci.*, **17**, 1893–1912, doi:10.5194/hess-17-1893-2013.

FAO/IIASA/ISRIC/ISSCAS/JRC, 2012: Harmonized World Soil Database (version 1.2). FAO and IIASA, accessed 9 October 2013. [Available online at http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/index.html?sb=1.]

Forman, B. A., R. H. Reichle, and M. Rodell, 2012: Assimilation of terrestrial water storage from GRACE in a snow-dominated basin. *Water Resour. Res.*, **48**, W01507, doi:10.1029/2011WR011239.

Gentine, P., T. J. Troy, B. R. Lintner, and K. L. Findell, 2012: Scaling in surface hydrology: Progress and challenges. *J. Contemp. Water Res. Educ.*, **147**, 28–40, doi:10.1111/j.1936-704X.2012.03105.x.

Giuntoli, I., B. Renard, J.-P. Vidal, and A. Bard, 2013: Low flows in France and their relationship to large-scale climate indices. *J. Hydrol.*, **482**, 105–118, doi:10.1016/j.jhydrol.2012.12.038.

Göckede, M., and Coauthors, 2008: Quality control of CarboEurope flux data—Part 1: Coupling footprint analyses with flux data quality assessment to evaluate sites in forest ecosystems. *Biogeosciences*, **5**, 433–450, doi:10.5194/bg-5-433-2008.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, **377**, 80–91, doi:10.1016/j.jhydrol.2009.08.003.

——, M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye, 2012: Towards a comprehensive assessment of model structural adequacy. *Water Resour. Res.*, **48**, W08301, doi:10.1029/2011WR011044.

——, C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian, 2014: Large-sample hydrology: A need to balance depth with breadth. *Hydrol. Earth Syst. Sci.*, **18**, 463–477, doi:10.5194/hess-18-463-2014.

Haddeland, I., B. V. Matheussen, and D. P. Lettenmaier, 2002: Influence of spatial resolution on simulated streamflow in a macroscale hydrologic model. *Water Resour. Res.*, **38**, 1124, doi:10.1029/2001WR000854.

Hargreaves, G., and Z. Samani, 1982: Estimating potential evapotranspiration. *J. Irrig. Drain. Div.*, **108**, 225–230.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.*, **113**, D20119, doi:10.1029/2008JD010201.

Hofstra, N., M. Haylock, M. New, and P. D. Jones, 2009: Testing E-OBS European high-resolution gridded data set of daily precipitation and surface temperature. *J. Geophys. Res.*, **114**, D21101, doi:10.1029/2009JD011799.

Horton, R., 1935: *Surface Runoff Phenomena. Part 1: Analysis of the Hydrograph.* Horton Hydrologic Laboratory Publication 101, Edward Bros, 73 pp.

Hundecha, Y., and A. Bárdossy, 2004: Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *J. Hydrol.*, **292**, 281–295, doi:10.1016/j.jhydrol.2004.01.002.

Hurst, H. E., 1951: Long-term storage of reservoirs: An experimental study. *Trans. Amer. Soc. Civ. Eng.*, **116**, 770–799.

Jacob, T., J. Wahr, W. T. Pfeffer, and S. Swenson, 2012: Recent contributions of glaciers and ice caps to sea level rise. *Nature*, **482**, 514–518, doi:10.1038/nature10847.

Jarvis, A., H. Reuter, A. Nelson, and E. Guevara, 2008: Hole-filled SRTM for the globe version 4. CGIAR-CSI SRTM 90m Database, CGIAR-CSI, accessed 18 March 2014. [Available online at http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database-v4-1.]

Jung, M., and Coauthors, 2011: Global patterns of land–atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res.*, **116**, G00J07, doi:10.1029/2010JG001566.

Kessomkiat, W., H.-J. H. Franssen, A. Graf, and H. Vereecken, 2013: Estimating random errors of eddy covariance data: An extended two-tower approach. *Agric. For. Meteor.*, **171–172**, 203–219, doi:10.1016/j.agrformet.2012.11.019.

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

Köhli, M., M. Schrön, M. Zreda, U. Schmidt, P. Dietrich, and S. Zacharias, 2015: Footprint characteristics revised for field-scale soil moisture monitoring with cosmic-ray neutrons. *Water Resour. Res.*, **51**, 5772–5790, doi:10.1002/2015WR017169.

Kuichling, E., 1889: The relation between the rainfall and the discharge of sewers in populous districts. *Trans. Amer. Soc. Civ. Eng.*, **20**, 1–56.

Kumar, R., L. Samaniego, and S. Attinger, 2010: The effects of spatial discretization and model parameterization on the prediction of extreme runoff characteristics. *J. Hydrol.*, **392**, 54–69, doi:10.1016/j.jhydrol.2010.07.047.

——, B. Livneh, and L. Samaniego, 2013a: Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme. *Water Resour. Res.*, **49**, 5700–5714, doi:10.1002/wrcr.20431.

——, L. Samaniego, and S. Attinger, 2013b: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resour. Res.*, **49**, 360–379, doi:10.1029/2012WR012195.

Landerer, F. W., and S. C. Swenson, 2012: Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resour. Res.*, **48**, W04531, doi:10.1029/2011WR011453.

Li, B., M. Rodell, B. F. Zaitchik, R. H. Reichle, R. D. Koster, and T. M. van Dam, 2012: Assimilation of GRACE terrestrial water storage into a land surface model: Evaluation and potential value for drought monitoring in western and

central Europe. *J. Hydrol.*, **446–447**, 103–115, doi:10.1016/j.jhydrol.2012.04.035.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, 1994: A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.*, **99**, 14 415–14 428, doi:10.1029/94JD00483.

Lievens, H., and Coauthors, 2015: Optimization of a radiative transfer forward operator for simulating SMOS brightness temperatures over the upper Mississippi basin. *J. Hydrometeor.*, **16**, 1109–1134, doi:10.1175/JHM-D-14-0052.1.

Liu, Y. Y., R. M. Parinussa, W. A. Dorigo, R. A. M. De Jeu, W. Wagner, A. I. J. M. van Dijk, M. F. McCabe, and J. P. Evans, 2011: Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals. *Hydrol. Earth Syst. Sci.*, **15**, 425–436, doi:10.5194/hess-15-425-2011.

Livneh, B., and D. Lettenmaier, 2012: Multi-criteria parameter estimation for the unified land model. *Hydrol. Earth Syst. Sci.*, **16**, 3029–3048, doi:10.5194/hess-16-3029-2012.

——, R. Kumar, and L. Samaniego, 2015: Influence of soil textural properties on hydrologic fluxes in the Mississippi River basin. *Hydrol. Processes*, **29**, 4638–4655, doi:10.1002/hyp.10601.

Lorenzo-Lacruz, J., S. Vicente-Serrano, J. López-Moreno, E. Morán-Tejeda, and J. Zabalza, 2012: Recent trends in Iberian streamflows (1945–2005). *J. Hydrol.*, **414–415**, 463–475, doi:10.1016/j.jhydrol.2011.11.023.

Mauder, M., T. Foken, R. Clement, J. A. Elbers, W. Eugster, T. Grünwald, B. Heusinkveld, and O. Kolle, 2008: Quality control of CarboEurope flux—Part 2: Inter-comparison of eddy-covariance software. *Biogeosciences*, **5**, 451–462, doi:10.5194/bg-5-451-2008.

Merz, R., J. Parajka, and G. Blöschl, 2009: Scale effects in conceptual hydrological modeling. *Water Resour. Res.*, **45**, W09405, doi:10.1029/2009WR007872.

Morris, M. D., 1991: Factorial sampling plans for preliminary computational experiments. *Technometrics*, **33**, 161–174, doi:10.1080/00401706.1991.10484804.

NASA, 2015: GRACE monthly mass grids—Land. Jet Propulsion Laboratory, accessed 1 July 2015. [Available online at http://grace.jpl.nasa.gov/data/get-data/monthly-mass-grids-land/.]

Nash, J. E., 1958: The form of the instantaneous unit hydrograph. *IAHS Publ.*, **45**, 114–121. [Available online at http://iahs.info/uploads/dms/045011.pdf.]

Nelsen, R. B., 2006: *An Introduction to Copulas.* Springer-Verlag, 269 pp., doi:10.1007/0-387-28678-0.

Oki, T., and S. Kanae, 2006: Global hydrological cycles and world water resources. *Science*, **313**, 1068–1072, doi:10.1126/science.1128845.

Orth, R., and S. I. Seneviratne, 2015: Introduction of a simple-model-based land surface dataset for Europe. *Environ. Res. Lett.*, **10**, 044012, doi:10.1088/1748-9326/10/4/044012.

Papale, D., and Coauthors, 2006: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: Algorithms and uncertainty estimation. *Biogeosciences*, **3**, 571–583, doi:10.5194/bg-3-571-2006.

Pokhrel, P., K. K. Yilmaz, and H. V. Gupta, 2012: Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *J. Hydrol.*, **418–419**, 49–60, doi:10.1016/j.jhydrol.2008.12.004.

Reed, S., V. Koren, M. Smith, Z. Zhang, F. Moreda, D.-J. Seo, and DMIP Participants, 2004: Overall distributed model intercomparison project results. *J. Hydrol.*, **298**, 27–60, doi:10.1016/j.jhydrol.2004.03.031.

Reichle, R. H., and R. D. Koster, 2004: Bias reduction in short records of satellite soil moisture. *Geophys. Res. Lett.*, **31**, L19501, doi:10.1029/2004GL020938.

Reichstein, M., and Coauthors, 2005: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Global Change Biol.*, **11**, 1424–1439, doi:10.1111/j.1365-2486.2005.001002.x.

Rodriguez-Iturbe, I., and J. B. Valdes, 1979: The geomorphologic structure of hydrologic response. *Water Resour. Res.*, **15**, 1409–1420, doi:10.1029/WR015i006p01409.

Rubel, F., and M. Kottek, 2010: Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen–Geiger climate classification. *Meteor. Z.*, **19**, 135–141, doi:10.1127/0941-2948/2010/0430.

Sakumura, C., S. Bettadpur, and S. Bruinsma, 2014: Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models. *Geophys. Res. Lett.*, **41**, 1389–1397, doi:10.1002/2013GL058632.

Samaniego, L., and A. Bárdossy, 2007: Relating macroclimatic circulation patterns with characteristics of floods and droughts at the mesoscale. *J. Hydrol.*, **335**, 109–123, doi:10.1016/j.jhydrol.2006.11.004.

——, R. Kumar, and S. Attinger, 2010: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.*, **46**, W05523, doi:10.1029/2008WR007327.

——, ——, and C. Jackisch, 2011: Predictions in a data-sparse region using a regionalized grid-based hydrologic model driven by remotely sensed data. *Hydrol. Res.*, **42**, 338–355, doi:10.2166/nh.2011.156.

——, ——, and M. Zink, 2013: Implications of parameter uncertainty on soil moisture drought analysis in Germany. *J. Hydrometeor.*, **14**, 47–68, doi:10.1175/JHM-D-12-075.1.

Sankarasubramanian, A., and R. M. Vogel, 2002: Comment on the paper: Basin hydrologic response relations to distributed physiographic descriptors and climate by Karen Plaut Berger, Dara Entekhabi, 2001. Journal of Hydrology 247, 169–182. *J. Hydrol.*, **263**, 257–261, doi:10.1016/S0022-1694(02)00061-6.

SCS, 1973: *A Method for Estimating Volume and Rate of Runoff in Small Watersheds.* SCS-TP-149, Soil Conservation Service, U.S. Department of Agriculture, 62 pp. [Available online at ftp://ftp.wcc.nrcs.usda.gov/wntsc/H&H/TRsTPs/TP149.pdf.]

Seibert, J., 2000: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrol. Earth Syst. Sci.*, **4**, 215–224, doi:10.5194/hess-4-215-2000.

Sherman, L. K., 1932: Streamflow from rainfall by the unit-graph method. *Eng. News Rec.*, **108**, 501–505.

Shuttleworth, W. J., 2012: *Terrestrial Hydrometeorology.* Wiley, 448 pp., doi:10.1002/9781119951933.

Sorooshian, S., and J. A. Dracup, 1980: Stochastic parameter estimation procedures for hydrologie rainfall–runoff models: Correlated and heteroscedastic error cases. *Water Resour. Res.*, **16**, 430–442, doi:10.1029/WR016i002p00430.

Stöckli, R., P. L. Vidale, A. Boone, and C. Schär, 2007: Impact of scale and aggregation on the terrestrial water exchange: Integrating land surface models and Rhône catchment observations. *J. Hydrometeor.*, **8**, 1002–1015, doi:10.1175/JHM613.1.

Su, H., Z.-L. Yang, R. E. Dickinson, C. R. Wilson, and G.-Y. Niu, 2010: Multisensor snow data assimilation at the continental scale: The value of Gravity Recovery and Climate Experiment terrestrial water storage information. *J. Geophys. Res.*, **115**, D10104, doi:10.1029/2009JD013035.

Sutanudjaja, E., L. Van Beek, S. De Jong, F. Van Geer, and M. Bierkens, 2014: Calibrating a large-extent high-resolution

coupled groundwater–land surface model using soil moisture and discharge data. *Water Resour. Res.*, **50**, 687–705, doi:10.1002/2013WR013807.

Swenson, S. C., and J. Wahr, 2006: Post-processing removal of correlated errors in GRACE data. *Geophys. Res. Lett.*, **33**, L08402, doi:10.1029/2005GL025285.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, doi:10.1029/2000JD900719.

Tetzlaff, D., S. K. Carey, H. Laudon, and K. McGuire, 2010: Catchment processes and heterogeneity at multiple scales—Benchmarking observations, conceptualization and prediction. *Hydrol. Processes*, **24**, 2203–2208, doi:10.1002/hyp.7784.

Thober, S., R. Kumar, J. Sheffield, J. Mai, D. Schaefer, and L. Samaniego, 2015: Seasonal soil moisture drought prediction over Europe using the North American Multi-Model Ensemble (NMME). *J. Hydrometeor.*, doi:10.1175/JHM-D-15-0053.1, in press.

Troy, T. J., E. F. Wood, and J. Sheffield, 2008: An efficient calibration method for continental-scale land surface modeling. *Water Resour. Res.*, **44**, W09411, doi:10.1029/2007WR006513.

Velpuri, N., G. Senay, R. Singh, S. Bohms, and J. Verdin, 2013: A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sens. Environ.*, **139**, 35–49, doi:10.1016/j.rse.2013.07.013.

Wada, Y., L. P. H. van Beek, C. M. van Kempen, J. W. T. M. Reckman, S. Vasak, and M. F. P. Bierkens, 2010: Global depletion of groundwater resources. *Geophys. Res. Lett.*, **37**, L20402, doi:10.1029/2010GL044571.

Wanders, N., D. Karssenberg, A. de Roo, S. M. de Jong, and M. F. P. Bierkens, 2014: The suitability of remotely sensed soil moisture for improving operational flood forecasting. *Hydrol. Earth Syst. Sci.*, **18**, 2343–2357, doi:10.5194/hess-18-2343-2014.

Werth, S., and A. Güntner, 2010: Calibration analysis for water storage variability of the global hydrological model WGHM. *Hydrol. Earth Syst. Sci.*, **14**, 59–78, doi:10.5194/hess-14-59-2010.

Wood, E. F., 1995: Scaling behaviour of hydrological fluxes and variables: Empirical studies using a hydrological model and remote sensing data. *Hydrol. Processes*, **9**, 331–346, doi:10.1002/hyp.3360090308.

Xia, Y., J. Sheffield, M. B. Ek, J. Dong, N. Chaney, H. Wei, J. Meng, and E. F. Wood, 2014: Evaluation of multi-model simulated soil moisture in NLDAS-2. *J. Hydrol.*, **512**, 107–125, doi:10.1016/j.jhydrol.2014.02.027.

——, M. T. Hobbins, Q. Mu, and M. B. Ek, 2015: Evaluation of NLDAS-2 evapotranspiration against tower flux site observations. *Hydrol. Processes*, **29**, 1757–1771, doi:10.1002/hyp.10299.

Yilmaz, K. K., H. V. Gupta, and T. Wagener, 2008: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resour. Res.*, **44**, W09417, doi:10.1029/2007WR006716.

Zaitchik, B. F., M. Rodell, and R. H. Reichle, 2008: Assimilation of GRACE terrestrial water storage data into a land surface model: Results for the Mississippi River basin. *J. Hydrometeor.*, **9**, 535–548, doi:10.1175/2007JHM951.1.

Zreda, M., W. Shuttleworth, X. Zeng, C. Zweck, D. Desilets, T. Franz, and R. Rosolem, 2012: COSMOS: The cosmic-ray soil moisture observing system. *Hydrol. Earth Syst. Sci.*, **16**, 4079–4099, doi:10.5194/hess-16-4079-2012.