JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001657

Key Points:
- Observation-based products of evapotranspiration and precipitation exhibit a large uncertainty at the watershed scale
- Model biases at the watershed scale are subject to observational uncertainty
- Incorporation of human water use improves the simulation of evapotranspiration and precipitation at the local scale

Supporting Information:
- Supporting Information S1

Correspondence to:
J. Keune,
jessica.keune@ugent.be

Citation:
Keune, J., Sulis, M., & Kollet, S. J. (2019). Potential added value of incorporating human water use on the simulation of evapotranspiration and precipitation in a continental-scale bedrock-to-atmosphere modeling system: A validation study considering observational uncertainty. *Journal of Advances in Modeling Earth Systems*, *11*, 1959–1980. https://doi.org/10.1029/2019MS001657

Received 12 FEB 2019
Accepted 17 MAY 2019
Accepted article online 28 MAY 2019
Published online 2 JUL 2019

Potential Added Value of Incorporating Human Water Use on the Simulation of Evapotranspiration and Precipitation in a Continental-Scale Bedrock-to-Atmosphere Modeling System: A Validation Study Considering Observational Uncertainty

J. Keune[1,2], M. Sulis[3], and S. J. Kollet[4,5]

[1]Meteorological Institute, University of Bonn, Bonn, Germany, [2]Laboratory of Hydrology and Water Management, Ghent University, Ghent, Belgium, [3]Environmental Research and Innovation, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg, [4]Institute for Bio- and Geosciences, Agrosphere (IBG-3), Research Centre Jülich, Jülich, Germany, [5]Centre for High-Performance Scientific Computing in Terrestrial Systems, Geoverbund ABC/J, Jülich, Germany

Abstract Human activities, such as human water use, have been shown to directly influence terrestrial water fluxes and states. Simulations of soil moisture, river discharge, evapotranspiration, and groundwater storage are significantly improved, if human interactions, such as irrigation and groundwater abstraction, are incorporated. Yet improvements through the incorporation of human water use on the simulation of local and remote precipitation are rarely studied but may contribute to the skill of land surface fluxes. In this study, we evaluate the impact of human water use on the skill of evapotranspiration and precipitation in a fully coupled bedrock-to-atmosphere modeling platform. The results show that human water use can potentially increase the skill of the simulations across scales. However, observational uncertainty at the watershed scale limits the identification of model deficiencies and added value related to human water use. Locally, daily precipitation statistics potentially benefit from the incorporation of human water use. Although the incorporation of human water use does not remove the wet bias, it can increase the model skill.

Plain Language Summary Precipitation forecasts exhibit large uncertainties, arising from unknown initial conditions and states, and limitations of models to represent processes at various scales. In order to improve simulations of the terrestrial water cycle, studies seek to improve the realism of the applied models through, for example, increased spatial resolution or the incorporation of additional processes across the Earth system. Here we analyze how the incorporation of additional processes in a European modeling system affects the accuracy of the simulated water fluxes, that is, evapotranspiration and precipitation. In particular, we improve the representation of groundwater in a continental-scale, atmospheric model and include groundwater pumping and irrigation, as major components of human water use. Results indicate that large-scale averages of precipitation, for example, over watersheds, are not necessarily improved if human water use is considered. However, significance of this finding is difficult to establish, because observational data sets exhibit large uncertainties, arising from, for example, the lack of observations in space and time, and miscalibration of the measuring devices. However, our results indicate that the accuracy of simulated daily precipitation is locally improved, which suggests that the incorporation of human water use may increase the accuracy of precipitation forecasts and advance our understanding of water cycle processes.

1. Introduction

The ability to understand, monitor, and predict the terrestrial water cycle under current and global change remains a major scientific and socioeconomic challenge (Wagener et al., 2010; Wood et al., 2011). Climate change scenarios along with an expected intensification of the water cycle with more frequent droughts and floods (Huntington, 2006) add uncertainty to future scenarios of water storages and flows (Vörösmarty et al., 2000). Commonly, the intensification is often measured at the interface between the land surface and atmosphere through combined changes in evapotranspiration (ET) or precipitation (P) as an indicator for drying, which has been discussed controversially (Greve et al., 2014). While the assessment

of human impacts on the terrestrial water cycle remains challenging (Oki & Kanae, 2006), studies demonstrated that human water management may lead to an intensification of the terrestrial water cycle, especially during droughts (e.g., He et al., 2017; Wanders & Wada, 2015).

Several attempts across multiple disciplines have been made to improve the simulation of the terrestrial water cycle. From an atmospheric perspective, these range from increasing resolutions to improve the simulation of P, for example, from general circulation models to regional circulation models (di Luca et al., 2015; Rummukainen, 2016) toward convection permitting simulations (Hohenegger & Schär, 2007; Prein et al., 2013, 2015) to replace simplified parameterization schemes. Yet, especially regional circulation models with parameterized convection have been identified to exhibit a wet bias (e.g., Casanueva et al., 2016; Rojas et al., 2011), which additionally limits their ability to simulate extremes and heat waves (Vautard et al., 2013; Weisheimer et al., 2011). In this case, a scale separation in space is beneficial to identify the added value, which is often located at the meso-β scale (20–200 km; Feser, 2006).

An improved representation of land surface states can further improve the realism of the simulated terrestrial water cycle. Here the incorporation of uncertainties in land surface models also leads to an improved predictability of P (Orth et al., 2016) and extremes, such as the heat wave over Europe in 2003 (MacLeod et al., 2016). This in turn might indicate the lack of information and realism that current land surface models comprise (Davin et al., 2016; Miralles et al., 2018; Tuinenburg & de Vries, 2017). Here even observation-based estimates of terrestrial ET exhibit relatively large uncertainties (Müller et al., 2011, 2013; Sörensson & Ruscica, 2018; Trambauer et al., 2014). Therefore, terrestrial system models and data sets to inform these models have been continuously expanded, incorporating additional processes of the water, energy, and biogeochemical cycles including human activities under the assumption that an increasing number and realism of simulated processes of the terrestrial water cycle leads to improved predictive skill (Sulis et al., 2018). One such process is the incorporation of human management of land cover, land use, and water resources. The most commonly studied management practice is irrigation, which may significantly influence P (DeAngelis et al., 2010; de Vrese et al., 2016), and its incorporation can improve the skill of general circulation model simulations (Thiery et al., 2017).

Large uncertainty in the simulation of the terrestrial water cycle also arises from the subsurface state. From the hydrologic perspective, the simulation of the terrestrial water cycle can be improved by increasing resolution (Bierkens et al., 2015) and the incorporation of human water use (HWU) beyond irrigation (Wada et al., 2017). Hydrologic model validation comprises a wide range of studies, which evaluate the performance of the simulated terrestrial water states and fluxes, such as soil moisture, ET, and discharge, from watersheds to continents (e.g., Döll et al., 2003; Haddeland et al., 2006; Liu et al., 2017; Rakovec et al., 2016). While the incorporation of water management, including groundwater abstraction and reservoirs, has been shown to improve the skill of simulating hydrological extremes (Veldkamp et al., 2018), a high uncertainty of water fluxes simulated with land surface and hydrological models comes from the driving atmospheric forcing and P (e.g., Biemans et al., 2009; Müller-Schmied et al., 2014, 2016; Wang et al., 2016).

In an effort to move toward a process-based and potentially more realistic model based representation of the terrestrial water cycle, studies attempt to bridge the gap between hydrologic and atmospheric research through the incorporation of a better representation of subsurface and land surface hydrology including a HWU interface in Earth System Models (Hazenberg et al., 2016; McDermid et al., 2017; Nazemi & Wheater, 2015; Pokhrel et al., 2015; Voisin et al., 2013). While progress has been made in understanding the terrestrial water cycle, the validation of such integrated modeling systems and the identification of the origin of biases become increasingly challenging (Oreskes et al., 1994) due to the lack of continuous and colocated observations across the Earth System from groundwater across the land surface into the atmosphere and observational uncertainty (Kotlarski et al., 2017). However, especially for water resource assessments, the identification of biases is essential. In a coupled modeling system, this bias might propagate not only from the local to the regional scale (Addor et al., 2016) but also from one compartment to the other (groundwater to land surface, land surface to atmosphere, and vice versa; Sulis et al., 2017), which potentially aggravates the bias through feedback processes.

There is a growing number of observations including satellites but also sensor networks and single stations to measure states and fluxes of the terrestrial water cycle (McCabe et al., 2017). Yet validation studies often neglect the fact that observations include additional uncertainty. There exist multiple sources that introduce

uncertainty to observations. For example, observational devices may exhibit malfunctions, miscalibration, and location biases. Analogously, all observations have a footprint, that is, a spatiotemporal reference, for which they are assumed to be representative. The footprint of point measurements is typically very small and on the order of a few meters, especially with respect to small-scale processes, such as ET, which is mainly influenced by local features, and convective P. On the contrary, satellite images have a much larger footprint on the order of tens to hundreds of kilometers. To bridge the gap between these scales, point measurements are often aggregated to gridded observational products, which introduce an additional uncertainty through assumptions of the underlying algorithm and are especially critical for regions with a sparse coverage of in situ stations.

In this study, we evaluate the integration of HWU in simulations using the continental scale integrated Terrestrial Systems Modeling Platform (TerrSysMP), a fully coupled soil-vegetation-atmosphere modeling system, during the heatwave 2003 in Europe. The main objectives of this study are to (1) evaluate the accuracy of integrated groundwater-atmosphere simulations at the land-atmosphere interface considering observational uncertainty and (2) assess potential added values of incorporating HWU on simulation skills with a focus on ET and P. The first objective aims at a general evaluation of the agreement and the accuracy of the simulations at the watershed scale at monthly to annual time scales using gridded observational data sets of ET and P. Multiple observational data sets of ET and P are used for validation. Both variables are evaluated separately but also combined to assess the accuracy and uncertainty of the net water flux at the land-atmosphere interface. The second objective focuses on the added value of HWU in the simulation of terrestrial water fluxes addressing the questions whether and how the consideration of HWU improves the skill of integrated bedrock-to-atmosphere simulations. Here additional daily in situ measurements of P and ET from colocated stations are utilized to discuss potential benefits of incorporating HWU.

The outline of this manuscript is as follows. Section 2 provides the methodology, the simulations, and the experimental design and introduces the observational data sets and metrics used to evaluate the accuracy and the skill of the simulations. Section 3 illustrates and discusses the results for a natural reference simulation. The added value of incorporating HWU is assessed in section 4. Summary and conclusions are provided in section 5.

2. Materials and Methods

2.1. Simulations

The simulations evaluated in this study were carried out with TerrSysMP (Shrestha et al., 2014; see also Gasper et al., 2014), which was setup over the European CORDEX domain at 0.11° resolution (Keune et al., 2016). HWU was considered as groundwater abstraction and irrigation (Keune et al., 2018). A detailed description of the setup can be found in aforementioned references, but a brief description is repeated here.

2.1.1. Setup of the Modeling System

TerrSysMP simulates the full terrestrial hydrologic cycle from groundwater to the atmosphere including feedbacks between shallow groundwater, land surface ET, and boundary layer processes along with P initiation. Lateral 3-D surface-subsurface flow is simulated using ParFlow (Jones & Woodward, 2001; Kollet & Maxwell, 2006; Maxwell, 2013), which is coupled to the land surface model CLM3.5 (Oleson et al., 2008) and the atmospheric weather prediction model COSMO (version 5.1; Baldauf et al., 2011; Doms & Schättler, 2002) through the Ocean Atmosphere Sea Ice Soil coupling tool OASIS3-MCT (Valcke, 2013). In this setup, COSMO was run with a 60-s time step, while ParFlow and CLM were run with a 180-s time step. Coupling between the models was performed every 180 s, constituting a very high coupling frequency to account for nonlinearities.

COSMO was initialized and driven at the boundaries with the reanalysis of the European Centre for Medium-Range Weather Forecasts, ERA-Interim (Dee et al., 2011). In order to keep the large-scale atmospheric circulation of all simulations consistent with the reanalysis, spectral nudging (von Storch et al., 2000) was applied. Vertical winds (u,v) were nudged above the planetary boundary layer ($p < 850$-hPa) with a nudging coefficient of $a = 0.05$ for wave numbers smaller than 14. The δ-two-stream approximation of the radiative transfer equation according to Ritter and Geleyn (1992) was used. Convection was parameterized

with the Tiedtke (1989) mass flux scheme, and vertical turbulent diffusion was simulated with a 2.5-level closure scheme. Cloud water, cloud ice, rain, and snow were simulated with a bulk-water continuity model.

The land surface characteristics for CLM3.5 were based on previous studies (Keune et al., 2016) including an updated land cover and leaf area index (LAI) data set based on the Moderate Resolution Imaging Spectroradiometer (Friedl et al., 2002). Subgrid-scale heterogeneity was accounted for by a maximum of four plant functional types per grid cell.

ParFlow was set up with the hydrofacies distribution 2 from Keune et al. (2016) in order to represent vertically heterogeneous soil and hydrogeologic characteristics. This hydrofacies distribution consists of a soil section (reaching a depth of 3 m) and a deeper subsurface and bedrock hydrogeology (reaching a depth of 103 m). The soil texture was prescribed using the Food and Agriculture Organization (FAO, 1988) soil database for the upper 10 soil layers. Hydrogeology was prescribed at the lower five soil layers using the Gleeson database (Gleeson, Marklund, et al., 2011; Gleeson, Smith, et al., 2011). To mimic the ocean, a lateral Dirichlet boundary condition with a hydrostatic profile of a shallow water table was applied.

2.1.2. Water Use Scenarios

In this study, a set of five simulations, carried out over the heatwave year 2003, was analyzed. These include a natural reference simulation (NAT) and four HWU scenarios. The HWU scenarios were constructed as follows: Two observation-based HWU data sets and two water use schedules were applied. HWU1 describes the HWU, that is, irrigation and (total) groundwater abstraction, from Wada et al. (2012, 2016). HWU2 is based on HWU from Siebert et al. (2010) and Siebert and Döll (2010). In both cases, groundwater abstraction accounted for irrigation and domestic and industrial demand, where domestic and industrial demand from Wada et al. (2012, 2016) were added to groundwater abstraction estimates for irrigation from Siebert et al. (2010) and Siebert and Döll (2010) for consistency. Irrigation and pumping were applied simultaneously between 06:00 and 18:00 UTC (daytime, HWU1-1 and HWU2-1) or between 18:00 and 06:00 UTC (nighttime, HWU1-2 and HWU2-2). Both data sets were bilinearly interpolated to the 0.11° grid, and rates of each daily estimate were added to the top soil layer (irrigation) or subtracted from the bottom soil layer (groundwater abstraction/pumping) at every time step in ParFlow using source/sink terms. All simulations were started from a multiyear spinup over the year 2003 after which a dynamic equilibrium was reached.

In this study, the simulations evaluated include the four water use scenarios (HWU's) and the natural reference simulation (NAT). Note that originally, the water use ensemble was not created to represent real-world conditions but to induce a relatively large spread in order to assess the uncertainty of land-atmosphere feedback related to HWU (Keune et al., 2018). Nevertheless, these simulations are used to assess the accuracy of each water use scenario but also the overall accuracy of the water use ensemble mean (mean of all four HWU's) and test improvements against the natural reference simulation. The water use ensemble mean is calculated using equally weighted water use scenarios, that is,

$$HWU{-}ENS = \sum_{i=1}^{M} w_i \cdot HWU_i,$$

with $w_i = 1/4$ for all scenarios and $M = 4$. Thus, six simulations overall (four water use scenarios, one water use mean, and one natural reference run) are evaluated using observations.

2.1.3. Simulation Period

Simulations were performed for the heatwave year 2003, which was characterized by extreme dry and hot conditions over Europe (Schär & Jendritzky, 2004) due to numerous hydrometeorological factors (Black et al., 2004). An anomalously anticyclonic high over Europe from May to August leads to reduced clouds and P, which in turn lead to desiccating soils. These effects induced a pre-heatwave at the beginning of July and a mega heatwave in August. Land surface conditions played a crucial role during both heatwave periods and, due to the absence of sufficient soil moisture to meet evaporative demands, were characterized by a heat accumulation through sensible heat. The mega heatwave period was influenced by a steady anticyclonic high over France (FR), which promoted the dominance of the local heat balance, which, in turn, lead to an exacerbating soil moisture-temperature feedback loop and extreme temperature records (Miralles et al., 2014). The exceptional combination of all feedbacks makes the heatwave of 2003 difficult to simulate with current models (Weisheimer et al., 2011). In particular, Weisheimer et al. (2011) showed that it requires not only an improved land surface hydrology but also improved radiation and convection parameterizations

Table 1
Overview of the Gridded Observational Data Sets of P and ET Used in This Study

Observation	Description	Time scale	Spatial scale	Variable (unit)	References	Data access
Gridded precipitation data sets						
CRU v. 4.01	Interpolated gauge observations, not bias corrected (angular-distance weighting interpolation)	Monthly	0.5°	P (mm/month)	Harris et al. (2014)	http://browse.ceda.ac.uk/browse/badc/cru/data/cru_ts/cru_ts_4.00/
GPCC	Interpolated gauge observations, not bias corrected	Monthly	0.5°	P (mm/month)	Rudolf et al. (2005)	https://www.esrl.noaa.gov/psd/data/gridded/data.gpcc.html
UDelP	Interpolated gauge observations	Monthly	0.5°	P (cm/month)	Legates and Willmott (1990)	https://www.esrl.noaa.gov/psd/data/gridded/data.UDel_AirT_Precip.html
PREC/L (NOAA) v. 4.01	Interpolated gauge observations, not bias corrected	Monthly	0.5°	P (mm/month)	Chen et al. (2002)	https://www.esrl.noaa.gov/psd/data/gridded/data.precl.html
MSWEP v.1.2	Multisource weighted ensemble observations (satellites, gauges, and reanalysis)	Monthly (aggregated from 3 hr)	0.1°	P (mm/day)	Beck, Vergopolan, et al. (2017) and Beck, van Dijk, et al. (2017)	http://www.gloh2o.org/
Gridded evapotranspiration data sets						
FLUXNET-MTE	Statistical upscaling of flux observations using a Model-Tree-Ensemble approach	Monthly	0.5°	LE (MJ/m^2/day)	Jung et al. (2009, 2010, 2011)	https://www.bgc-jena.mpg.de/geodb/projects/Data.php
GLEAM v.3.1a	Priestley and Taylor based algorithm based on satellite observations	Monthly (aggregated from daily)	0.25°	ET (mm/day)	Miralles et al. (2011) and Martens et al. (2017)	http://www.gleam.eu/
CSIRO	Derived from the Budyko framework with observed P and Penman-Monteith E	Monthly	0.5°	ET (mm/month)	Leuning et al. (2008) and Zhang et al. (2016)	https://data.csiro.au/dap/landingpage?pid=csiro%3A17375

Note. P = precipitation; ET = evapotranspiration; LE = latent heat of evaporation.

to improve model skill in simulating the heatwave conditions. Here we assess how the integration of HWU improves the model skill.

2.2. Observations

Two types of observations are used for validation. First, gridded, model-based observations are used to evaluate monthly, seasonal, and annual estimates of ET, P, and the net moisture flux ET-P, at the watershed scale. All gridded observation-based data sets of ET and P that are available for 2003 at a spatial resolution of at least 0.5° and a monthly time scale are used for validation in order to assess observational uncertainty. These include only diagnostic data sets, which are either interpolated in situ observations, merged observational data sets, and in case of ET, are based on the Penman-Monteith equation or empirically derived formulations, such as the Priestley-Taylor equation. We refrain from using modeled estimates, in which ET and P are estimated with the help of a land surface model and/or a reanalysis, because these do not consider HWU explicitly (e.g., irrigation; Tuinenburg & de Vries, 2017). The gridded observations comprise five P and three ET data sets, which have been used in previous studies (Greve et al., 2014; Martens et al., 2018; Müller et al., 2011). Here we also combine these observations into 15 observational data sets of ET-P to assess the uncertainty of the observed net moisture flux at the watershed scale.

Second, colocated in situ measurements of ET and P from FLUXNET eddy-covariance sites are used to validate local-scale and daily variations of the net moisture flux and to identify the added value of incorporating HWU in fully integrated modeling systems at the local scale. In addition, a larger data set of in situ rain gauge measurements is used to identify skills at the local scale. Tables 1 and 2 provide an overview of the gridded and in situ observations used in this study, their provenance, and spatial and temporal resolution. A detailed description of each data set is provided below.

Table 2
Overview of the In Situ Observational Data Sets of P and ET Used in This Study

Observation	Description	Time scale	Spatial scale	Variable (unit)	References	Data access
In situ observations						
FLUXNET	Eddy-covariance measurements	Daily (aggregated from half hourly)	In situ	P (mm/day) and LE (W/m^2)	Baldocchi et al. (2001)	http://fluxnet.fluxdata.org/
ECA&D (blended)	Weather stations/rain gauges	Daily	In situ	P (mm/day)	Klein Tank et al. (2002)	http://www.ecad.eu/dailydata/predefinedseries.php

Note. P = precipitation; ET = evapotranspiration; LE = latent heat of evaporation.

2.2.1. Gridded Observations of ET and P

2.2.1.1. CRU

Climate Research Unit (CRU) TS4.01 (Harris et al., 2014) is a gridded time series data set, which was released in January 2017. It provides global, monthly observations of P at 0.5° × 0.5° resolution. Monthly station measurements are interpolated to a high-resolution grid by the CRU of the University of East Anglia (http://www.cru.uea.ac.uk/) to provide a global, homogenous climatic time series of basic variables.

2.2.1.2. GPCC

The global P data set from the Global Precipitation Climatology Centre (GPCC) is based on 67,200 quality-controlled gauge measurements and contains the monthly records of P interpolated to a regular 0.5° × 0.5° grid (Rudolf et al., 2005; Schneider et al., 2011). This gridded gauge-analysis product is not bias corrected and may include systematic gauge errors.

2.2.1.3. UDelP

The P data set from the University of Delaware (UDelP; Willmott & Matsuura, 2001; see also Legates & Willmott, 1990) combines gauge measurements of P from multiple climate monitoring networks, such as the Global Historical Climatology Network. The gauge observations were interpolated to a regular 0.5° × 0.5° grid using the Climatologically aided interpolation (Willmott & Robeson, 1995) and were not corrected for sensor biases.

2.2.1.4. PREC/L

The National Oceanic and Atmospheric Administration released a reconstructed P product over land (PREC/L) based on 17,000 gauge observations from the monitoring networks Global Historical Climatology Network and Climate Anomaly Monitoring System. Observations were interpolated to monthly averaged P over a regular 0.5° × 0.5° grid (Chen et al., 2002).

2.2.1.5. MSWEP

The multisource weighted ensemble precipitation (MSWEP) is a new P product, which was released in 2017 (Beck, Vergopolan, et al., 2017). It provides 3-hourly P at 0.1° resolution worldwide and combines gauge, satellite, and reanalysis data in a weighted ensemble approach. MSWEP adjusts for gauge biases on a daily basis and has been developed specifically for hydrological applications (Beck, van Dijk, et al., 2017).

2.2.1.6. FLUXNET-MTE

The model tree ensemble (MTE) proposed by Jung et al. (2009, 2010, 2011) is a machine learning approach to upscale eddy covariance measurements at FLUXNET stations (described below) to a global gridded data set. The machine learning technique was trained using remote sensing data sets, land use information, climate, and meteorological data (Jung et al., 2011). Thus, in regions of sparse FLUXNET measurements, the resulting fields rely strongly on the information content and accuracy of the training data set. The FLUXNET-MTE data set provides monthly averages of, for example, ET at 0.5° × 0.5° resolution and is often used as an independent data set for model evaluation and calibration (e.g., Bonan et al., 2011; Swenson & Lawrence, 2014).

2.2.1.7. GLEAM

The Global Land Evaporation Amsterdam Model (GLEAM; Miralles et al., 2011) estimates ET based on satellite observations. In GLEAM, ET is calculated as the residual of potential ET from the Priestley-Taylor equation constrained by water stress obtained from satellite observations of soil moisture and vegetation optical depth and conditioned on estimates and reanalyses of P, radiation, and temperature. In this

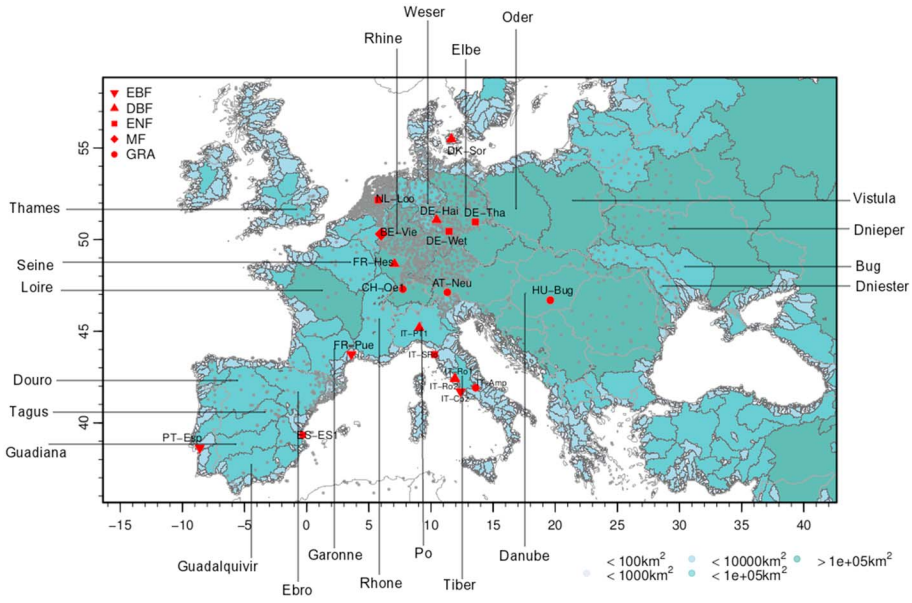


Figure 1. European watersheds larger than 100 km^2 considered in this study. Colors are representative of the watershed area. Watershed boundaries are taken from the Hydrological data and maps based on SHuttle Elevation Derivatives at multiple Scales (https://hydrosheds.cr.usgs.gov/, Lehner et al., 2006). Gray and red points indicate in situ observations from ECA&D and FLUXNET stations, respectively. The symbols are commensurate to the dominant land use of the FLUXNET station: evergreen broadleaf forest (EBF), deciduous broadleaf forest (DBF), evergreen needleleaf forest (ENF), mixed forest (MF), and grassland (GRA).

study, the GLEAM version 3.1a was used (Martens et al., 2017). GLEAM provides daily estimates of ET at 0.25° × 0.25° resolution, which have been shown to agree well with existing FLUXNET stations (Martens et al., 2017). GLEAM retrievals of root zone soil moisture and ET are often used in land-atmosphere feedback studies (e.g., Guillod et al., 2015; Miralles et al., 2012).

2.2.1.8. CSIRO

The ET data set from the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia provides global estimates of land surface ET and its components at 0.5° resolution. ET is derived in a Budyko framework using observed P and potential ET calculated with the Penman-Monteith-Leuning model (Leuning et al., 2008; Zhang et al., 2016).

2.2.2. In situ Observations of ET and P

2.2.2.1. FLUXNET

The European FLUXNET community (e.g., Baldocchi et al., 2001, http://fluxnet.fluxdata.org/) provides an error-corrected and quality-checked, combined data set of measurements across registered European eddy-covariance stations with half-hourly resolution. For 2003, a total of 19 stations over Central and Southern Europe is considered for validation, using a data coverage of at least 50% for P and ET, respectively. This selection comprises four eddy-covariance stations over grasslands, five stations over evergreen needleleaf forests, six stations over deciduous broadleaf forests, three stations over evergreen broadleaf forests, and one station over mixed forest. Three stations (i.e., IT-Col [deciduous broadleaf forests], NL-HAA [grasslands], and CZ-BK1 [evergreen needleleaf forests]) were neglected for analysis due to quality concerns. Figure 1 provides an overview of the FLUXNET stations used for validation. Note that uncorrected measurements of the latent heat flux are used for comparison and that eddy covariance stations do not necessarily close the land surface energy budget (e.g., Wilson et al., 2002).

2.2.2.2. ECA&D

The European Climate Assessment & Data (ECA&D, http://www.ecad.eu/) collection provides a comprehensive data set of daily in situ observations of multiple variables including P. Figure 1 shows the spatial distribution of P stations available for 2003 across Europe. Only stations with at least 95% data coverage (i.e., 342 days of 360 days excluding the model spinup) were selected for validation. This yields a total of 1,033 stations across Europe.

2.2.3. Data Aggregation

For all gridded observations at different spatial resolutions, watershed means are calculated based on the original resolution (native grid) and compared to the watershed average of the simulations. Daily values of gridded data sets, that is, GLEAM and MSWEP, are summarized to monthly values. The half-hourly values of FLUXNET are accumulated to daily latent heat fluxes (LE) and converted to daily ET (excluding missing values consistently from model results) and daily P. For the validation with station data, simulations are interpolated to the stations using the nearest neighbor method. For the comparison with FLUXNET stations, the nearest neighbor is used irrespective of the predominant land use type.

2.3. Validation Metrics and Skill Scores

The quality and accuracy of the simulation experiments is evaluated with observations at the monthly, seasonal, and annual watershed scale and at the daily and local scale. The following measures are used. The bias indicates the average direction of the error between simulation f and observation o, that is,

$$bias = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)$$

over a time series of length n. Here f and o are ET, P, or their difference, from simulation and observation, respectively. The bias provides only an average error between simulation and observation and does not indicate the quality of the simulations. The mean squared error (MSE) is used to measure the accuracy of the forecast, that is, the magnitude of the error weighted with the squares of error as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2.$$

The MSE skill score, SS_{MSE} (Jolliffe & Stephenson, 2012), relates the improvement of one simulation over another and allows to identify potential improvements of accuracy between the water use scenarios and the natural reference simulation. The SS_{MSE} is calculated as

$$SS_{\mathrm{MSE}} = \frac{MSE_{\mathrm{HWU}} - MSE_{\mathrm{NAT}}}{MSE_{\mathrm{perf}} - MSE_{\mathrm{NAT}}} = 1 - \frac{MSE_{\mathrm{HWU}}}{MSE_{\mathrm{NAT}}},$$

where MSE_{HWU} is the MSE of each water use scenario and MSE_{NAT} is the MSE of the natural reference simulation. MSE_{perf} describes the error of a perfect forecast with $MSE_{\mathrm{perf}} = 0$. Values close to 0 indicate no change of accuracy by incorporating HWU. A value of 1 indicates an improvement of 100%. The SS_{MSE} is a positively oriented score, that is, the higher the value, the stronger the improvement.

Similarly to Kotlarski et al. (2017), we consider observational uncertainty in our comparison. Here observational uncertainty is defined as the standard deviation of the variable's values among the observations.

Daily, in situ observations also allow to evaluate deterministic forecasts in a probabilistic approach, based on, for example, the joint occurrence of events. For P, we define thresholds for wet days (WD, P > 1 mm/day), heavy P days (HPD, P > 10 mm/day), and very heavy P days (VHPD, P > 20 mm/day), thus converting continuous values of P into dichotomous data sets of “yes”/“no” events. These events are then evaluated probabilistically by their joint occurrence in simulations and observations. Therefore, a contingency table (supporting information Table S1) is defined for each threshold, which allows to evaluate the conditional probability of an event being simulated, given the observations and vice versa. A number of skill scores have been identified based on the contingency table (e.g., Jolliffe & Stephenson, 2012), such as the frequency bias (FBI)

$$FBI = \frac{hits + false\quad alarms}{hits + misses},$$

which indicates the tendency of the simulations to overpredict or underpredict events. A value larger than 1 indicates that more events were simulated than observed. Vice versa, values below 1 indicate

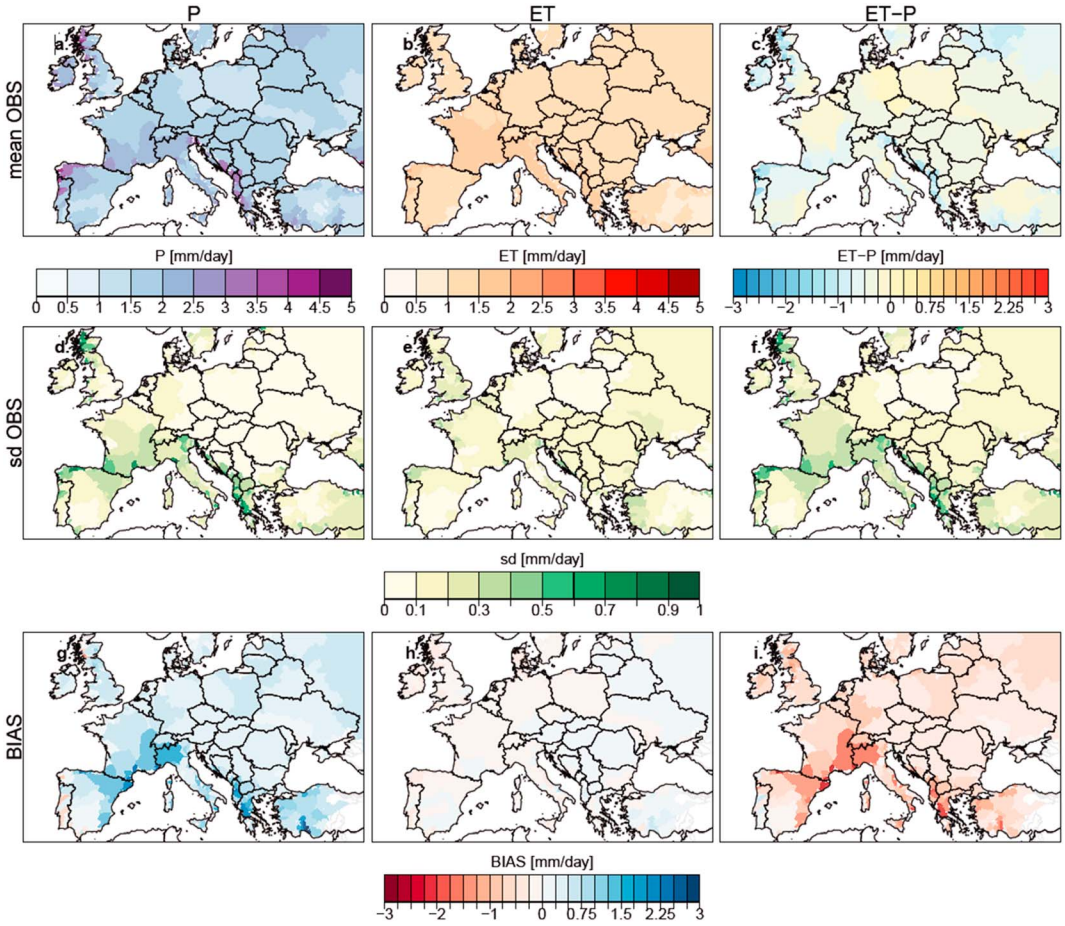


Figure 2. Maps of annually averaged P (mm/day), ET (mm/day), and ET-P (mm/day) over European watersheds in 2003. (a)–(c) and (d)–(f) show the mean and the standard deviation over the 5, 3, and 15 observations (each combination of ET: FLUXNET-MTE, CSIRO, and GLEAM with P: GPCC, CRU, PREC/L, MSWEP, and UDelP), respectively. (g)–(i) shows the mean bias of the NAT simulation of TerrSysMP with respect to the mean observations, again for P, ET, and ET-P. P = precipitation; ET = evapotranspiration.

underprediction. The Peirce's skill score (PSS) evaluates how well the simulations are able to separate the occurrence of events (yes events) from the nonoccurrence (no events). The PSS is defined as

$$PSS = \frac{hits}{hits + misses} - \frac{false\ \ alarms}{false\ \ alarms + correct\ \ negatives}$$

and hence evaluates the relation between the probability of detection (first term) and the probability of false detection (second term). The PSS varies between −1 and 1, where 0 indicates no skill. A perfect separation between occurrence and nonoccurrence of events is indicated by a PSS of 1.

3. Validation of the Natural Reference Simulation

This section evaluates the accuracy of the natural reference simulation with TerrSysMP over the heatwave year 2003. It is split in two parts: (i) the monthly watershed scale, at which observational uncertainty of gridded products is considered, and (ii) the daily, local scale using in situ observations.

3.1. Watershed Scale

In order to consider the observational uncertainty, we evaluate the accuracy of the simulated terrestrial water cycle based on multiple gridded observation-based products of ET and P at the watershed scale. Figure 2 shows ET, P, and ET-P averaged over the watersheds for the year 2003. The upper two rows show the average and the standard deviation over the 5, 3, and 15 observation-based products, respectively. The

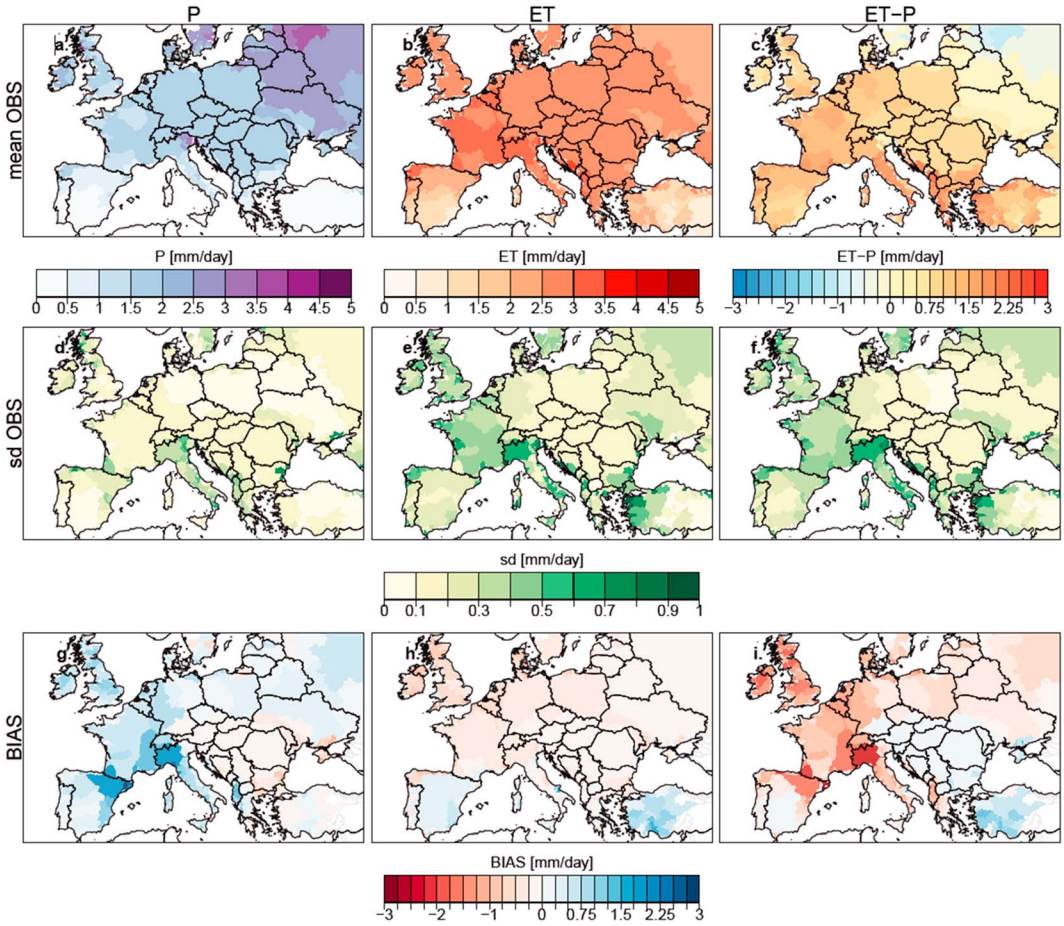


Figure 3. Same as Figure 2 but for summer 2003 (June-July-August).

bottom row shows the bias of the natural reference simulation with respect to the mean of all observations. Over the year, the average P commonly exceeds ET, but a few watersheds show a positive net moisture flux (e.g., Guadalquivir, Guadiana, Seine, Loire, Elbe, and Oder; Figure 2c). The model shows a positive P bias, especially over mountainous watersheds. The ET bias is comparably small, which leads to an underestimation of the net moisture flux by the model over almost all watersheds (Figure 2i). Note the highest observational uncertainties, arising through P, coincide with the highest biases (Figures 2d and 2g and 2f and 2i). Especially in mountainous regions, such as the Alps, the Pyrenees, the Cantabrian mountains (NE Spain), and the mountain ranges of the Iberian system (NW Spain), ET-P is underestimated, but the observations exhibit a high uncertainty.

During the dry summer (June, July, and August) of 2003, all European watersheds lost more water through ET than they received through P, which resulted in a positive upward moisture flux (Figures 3a–3c). Most watersheds over FR, the Iberian Peninsula (IB) and the Mediterranean (MD) received almost no P and indicate a net loss of water on the order of 2–3 mm/day. Whereas the model captures the net loss of water over the Southern European watersheds quite well, it simulates a negative ET-P balance over the watersheds Ebro, Rhone, and Po, resulting in a strong negative bias (Figure 3i), which also indicates the lack of skill in simulating heatwaves (Weisheimer et al., 2011). Again, this coincides with the highest observational uncertainty, determined by observational products of ET during summer (Figures 3e and 3f). Interestingly, the Po watershed, which is heavily water managed (Wada et al., 2016), shows the largest observational uncertainty of ET-P during summer.

Observational uncertainty is mainly determined by P over the course of the year and by ET during summer. Figure 4 shows exemplarily the annual cycle of ET, P, and ET-P for the Guadalquivir basin. The annual cycle of the water fluxes is well captured for this watershed but also indicates some differences between the model

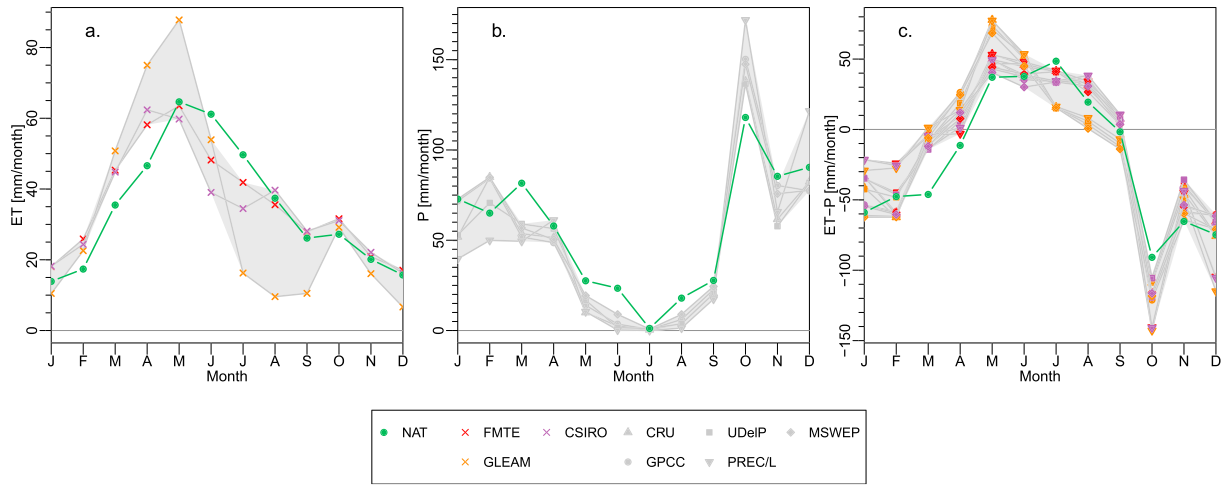


Figure 4. Monthly time series of (a) ET, (b) P, and (c) ET-P in (mm/month) averaged over the Guadalquivir watershed for all gridded observational data sets and the NAT simulation. The gray area indicates the maximum range covered by the observations. P = precipitation; ET = evapotranspiration.

and the observations. The observations show a peak of ET in late spring (Figure 4a), presumably through the prolonged P deficit and the pre-heatwave from March to May. However, the ET observations also show a large uncertainty from April to September, which is mainly stemming from GLEAM, and might indicate the strongest water limitation of the observational data sets from June onward. In contrast, this effect is hardly visible in case of CSIRO and FLUXNET-MTE, which agree better with the simulation results. The origin of the existing biases cannot be traced back conclusively to P, which is only marginally overestimated in this case. Vice versa, it remains unclear to which extent the ET observations are able to capture irrigation effects. In case of GLEAM, this is mainly limited to the capability of large-scale satellite measurements of soil moisture to capture the effect of (small-scale) irrigation (Lawston et al., 2017). Similarly, FLUXNET-MTE is limited to the availability of tower measurements, which might not be located in irrigated areas. Considering all these uncertainties, the annual cycle of ET-P is well captured, with an underestimation in spring and a shift of the peak from May to July.

3.2. Local Scale

The validation of daily ET, P, and ET-P fluxes with in situ measurements of 1,033 rain gauges and 19 FLUXNET towers confirms the above watershed-based biases. Figure 5 shows the annual (Figures 5a–5c) and summer (Figures 5d–5f) biases of P, ET, and ET-P for all available stations in 2003. Over the year, the simulation exhibits a positive P bias on the order of 0.5 mm/day based on FLUXNET stations and 0.49 mm/day based on ECA&D stations. This value is comparable to other validation studies (e.g., Katragkou et al., 2015; Kotlarski et al., 2014, 2017; Vautard et al., 2013). In contrast to the watershed scale, ET is overestimated, with a bias of 0.24 mm/day. For all available colocated measurements, this results in an underestimation of ET-P on the order of −0.29 mm/day. These biases are slightly increased for summer (0.94 mm/day [P, FLUXNET], 0.69 mm/day [P, ECA&D], 0.46 mm/day [ET, FLUXNET], and −0.48 mm/day [ET-P, FLUXNET]). Noteworthy is the high P bias over all stations in the area of Ebro and the Alps, which corroborates the well-known dependence on topography. On the other hand, the biases at FLUXNET stations reveal no dependence on the land use, which suggests the dominant role of P biases in the simulation. Yet it has to be mentioned that the lack of stations, and especially colocated stations, is a major limitation for this and many other studies.

A detailed analysis of the daily P statistics reveals that heavy and very heavy P events are most commonly responsible for an increased P bias. Figure 6 shows boxplots of the FBI and the PSS for WD, HPD, and VHPD, separately for the regions Mid-Europe (ME), FR, IB, MD, and Eastern Europe (EA). The FBI shows only a small overprediction of WD, with the highest skill in ME and FR. The FBI increases with the P threshold and shows an overprediction of HPD and VHPD. Over MD and IB, HPD (VHPD) are on average 2–3 (6–8) times more often simulated than observed. The skill also continuously decreases and indicates that over the depicted regions, the simulation is unable to correctly detect timing and location of VHPD.

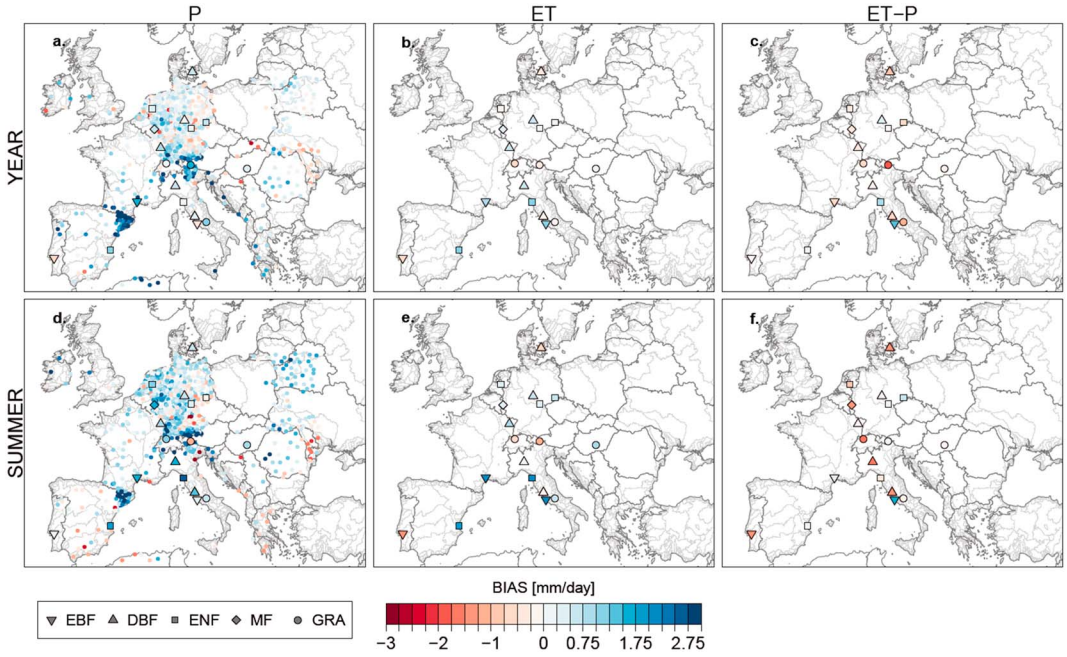


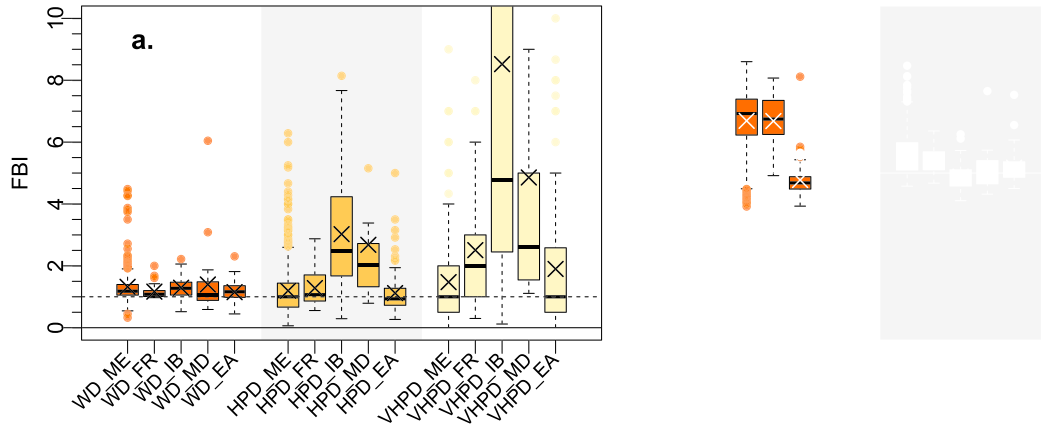
Figure 5. Daily P, ET, and ET-P biases from the NAT simulation of all available ECA&D and FLUXNET stations for (a–c) the full year 2003 and (d–f) summer 2003. ECA&D stations are illustrated by colored dots. FLUXNET stations are indicated by a black border around the symbols, which indicate the land use of the station. P = precipitation; ET = evapotranspiration; EBF = evergreen broadleaf forest; DBF = deciduous broadleaf forest; ENF = evergreen needleleaf forest; MF = mixed forest; GRA = grassland.

4. Added Value of Incorporating HWU

The question arises, to which extent the incorporation of HWU increases the skill of the simulations. Here the added value of incorporating HWU on (i) monthly fluxes at the watershed scale and (ii) daily, local fluxes of ET, P, and ET-P is assessed.

4.1. Watershed Scale

Figure 7a shows the annual bias of ET-P over 15 large watersheds across Europe as a boxplot for each simulation including all observation-based products. A comparison of the boxplots in each column (i.e., for each watershed) shows that the incorporation of HWU does not remove the overall bias (see also Figure 4). The



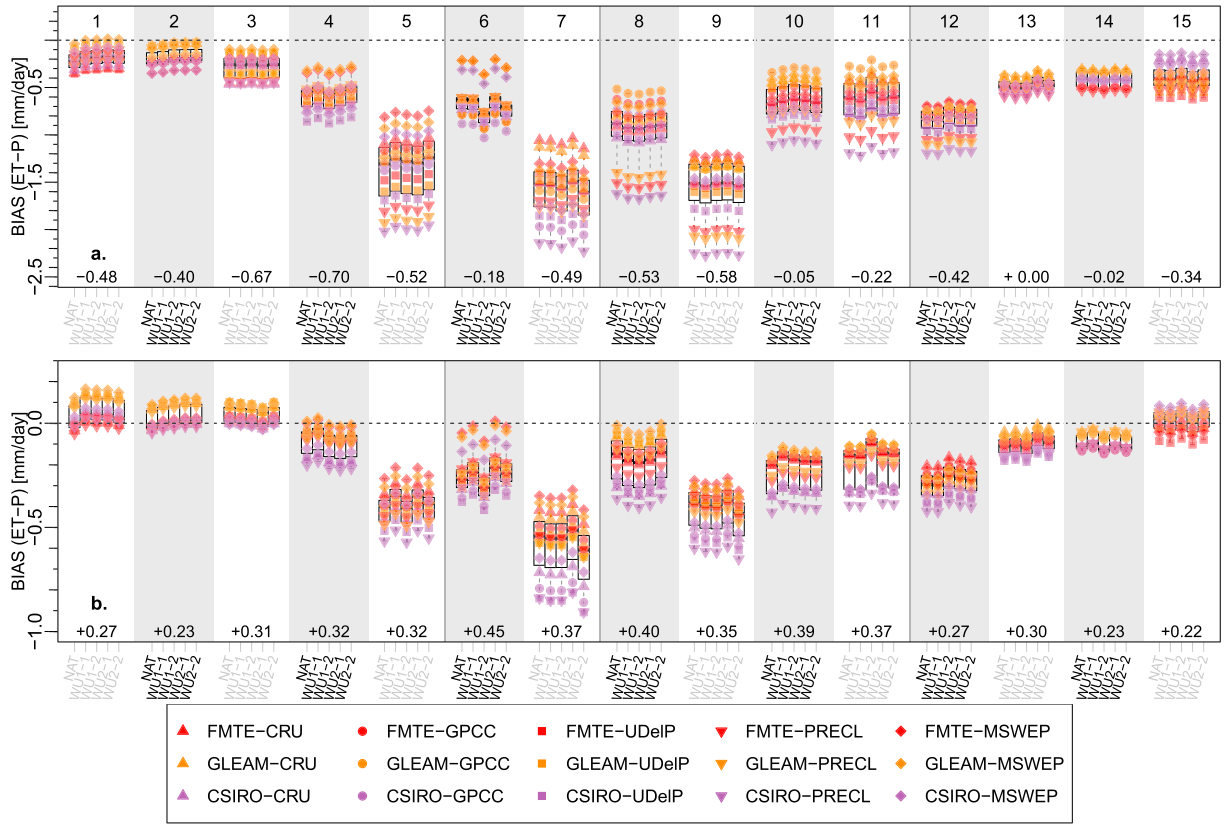


Figure 7. Box-Whisker Plots of the ET-P bias for (a) the full year 2003 and (b) summer 2003 for the 5 simulations, the 15 observations, and over 15 selected watersheds: 1 = Guadalquivir; 2 = Guadiana; 3 = Tagus; 4 = Douro; 5 = Ebro; 6 = Tiber; 7 = Po; 8 = Garonne; 9 = Rhone; 10 = Seine; 11 = Loire; 12 = Rhine; 13 = Elbe; 14 = Oder; 15 = Danube. The five boxplots for each watershed indicate the observational uncertainty of the simulated bias in the following order: NAT, HWU1-1, HWU1-2, HWU2-1, and HWU2-2. The symbols within each box indicate the observational P data set (Δ: CRU, ◯: GPCC; ▢: UDelP, ▽: PREC/L, and ◇: MSWEP) and the color the observational ET data set (red: FLUXNET-MTE, orange: GLEAM, and purple: CSIRO). A negative bias in (a) indicates that the net moisture flux ET-P is overestimated (simulations too wet) and in (b) that the continental source is underestimated. A positive bias in (b) indicates that ET-P is overestimated. The values at the bottom of each subplot indicate the mean ET-P flux of all observations over the watershed. P = precipitation; ET = evapotranspiration; HWU = human water use.

bias is rather subject to observational uncertainty, which increases from flat and coastal areas (Guadalquivir, Guadiana, and Tagus) toward mountainous watersheds (Douro, Ebro, Garonne, and Rhone). Here the observational uncertainty clearly exceeds the uncertainty related to HWU.

The high observational uncertainty of ET-P leads to a high uncertainty of the bias especially in summer (Figure 7b). As has been discussed before, the models fail to capture the prolonged drought and exhibit consistently a small overestimation of the net moisture flux especially for the southern watersheds over the IB. For most of the analyzed watersheds, the bias is negative in summer and hence indicates an underestimation of ET-P, either through an underestimation of ET or an overestimation of P (cf. Figure 3). The summer bias is the largest for the CSIRO data sets, ranging from, for example, −0.9 to −0.4 mm/day for the Po basin. On the contrary, the biases for Douro, Tiber, Garonne, and Elbe can be either negative or approximately 0 and depend on the observation-based products used. The fact that the highest biases coincide with the highest observational uncertainty, and furthermore exceed the variations of ET-P induced by HWU, hampers an objective validation. Moreover, this raises concerns especially for the use of gridded validation data sets in water assessment and scarcity studies.

The individual HWU scenarios were not constructed to represent real-world conditions but rather to assess local and nonlocal feedback processes in the terrestrial water and energy cycles associated with HWU (Keune et al., 2018). Hence, we do not expect the single HWU scenarios to improve model skill with respect to NAT. The mean of all scenarios, HWU-ENS, is assumed to be representative of the effect of a more

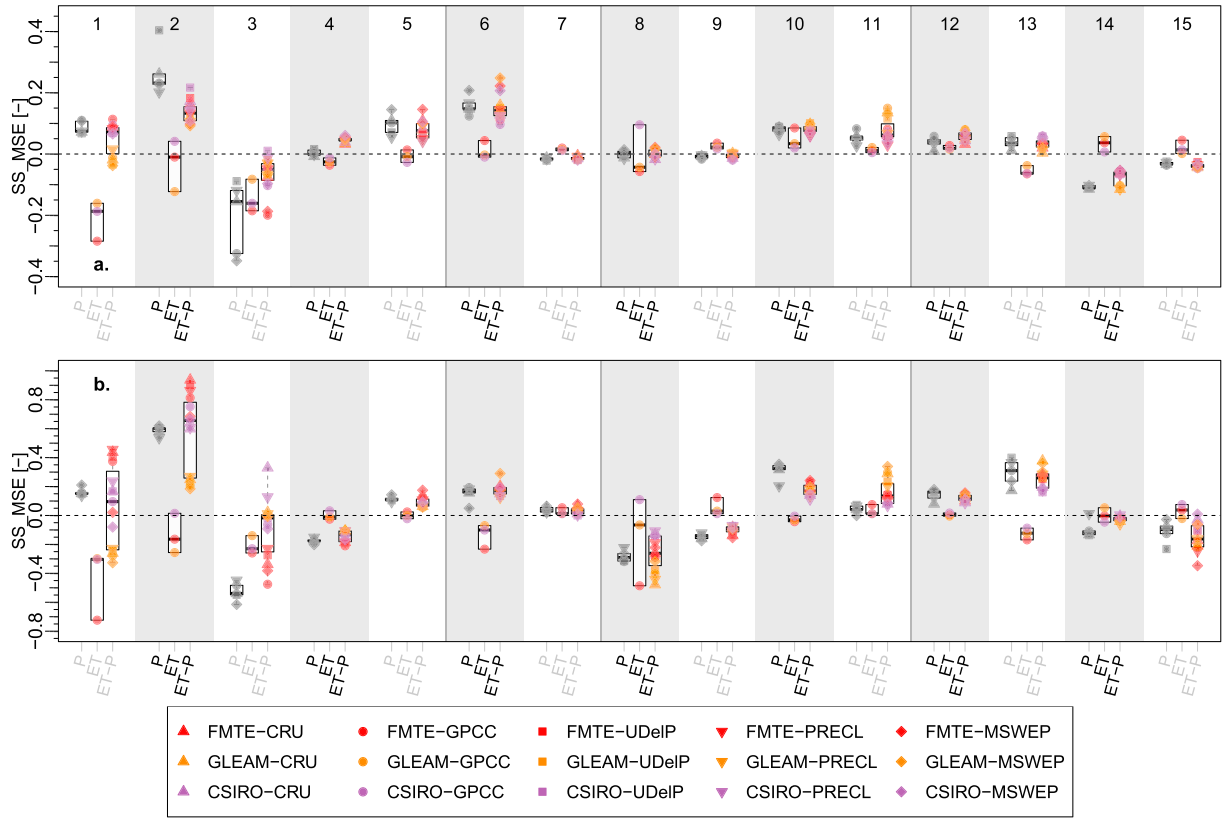


Figure 8. Box-Whisker Plots of the SS_{MSE} for P, ET, and ET-P of HWU-ENS over HWU-NAT, for (a) the full year 2003 and (b) summer 2003 over the 15 selected watersheds from Figure 7: 1 = Guadalquivir; 2 = Guadiana; 3 = Tagus; 4 = Douro; 5 = Ebro; 6 = Tiber; 7 = Po; 8 = Garonne; 9 = Rhone; 10 = Seine; 11 = Loire; 12 = Rhine; 13 = Elbe; 14 = Oder; 15 = Danube. The three boxplots for each watershed indicate the uncertainty of the SS_{MSE} with respect to the observational data set: P (Δ: CRU, ◯: GPCC; ▢: UDelP, ▽: PREC/L, and ◇: MSWEP), ET (red: FLUXNET-MTE, orange: GLEAM, and purple: CSIRO), and ET-P. A positive (negative) SS_{MSE} indicates that the incorporation of HWU increases (decreases) the skill of the simulations. A value of 0 indicates no change of skill.
P = precipitation; ET = evapotranspiration; HWU = human water use; MSE = mean squared error.

realistic representation of HWU. Hence, we focus on the added value of HWU-ENS over NAT in the following.

Figure 8 shows the SS_{MSE} of HWU-ENS over NAT and its dependency on observational uncertainty. The boxplots show the SS_{MSE} of P, ET, and ET-P for the 15 watersheds from Figure 7, for the full year (Figure 8a) and summer (Figure 8b), and include the 5, 3, and 15 observational products, respectively. Most watersheds indicate a clear improvement in the simulation of P and ET-P (e.g., Guadiana, Ebro, Tiber, and Elbe) but a deterioration of skill in the simulation of ET over the summer and the full year. The magnitude of this skill improvement for ET-P varies between 1% and 30% (average: 2.67% for the full year and 3.83% for the summer). However, for some watersheds, the added value is masked by observational uncertainty, either in magnitude (e.g., Guadiana) or in sign (e.g., Guadalquivir and Garonne).

Over all watersheds (including the ones not illustrated here), the SS_{MSE} indicates that the incorporation of HWU-ENS improves marginally the simulation of ET by 0.25% but deteriorates the skill for P and ET-P fluxes by some 1%. On the contrary, the simulation of the summer net moisture flux is improved by about 8%, which mainly stems from an improved simulation of P (~11%).

4.2. Local Scale

The differences associated with HWU increase toward the daily and local scales (Keune et al., 2018). To assess the added value of incorporating HWU at the local scale, daily in situ observations of ET, P, and ET-P are used. A SS_{MSE} is calculated for the 19 available FLUXNET stations. The results are shown in Figure 9, for ET-P, ET, and P along the rows and for the entire year and summer along the columns. Over the year and all available FLUXNET stations, the HWU mean improved the MSE of ET-P by ~13%, P by

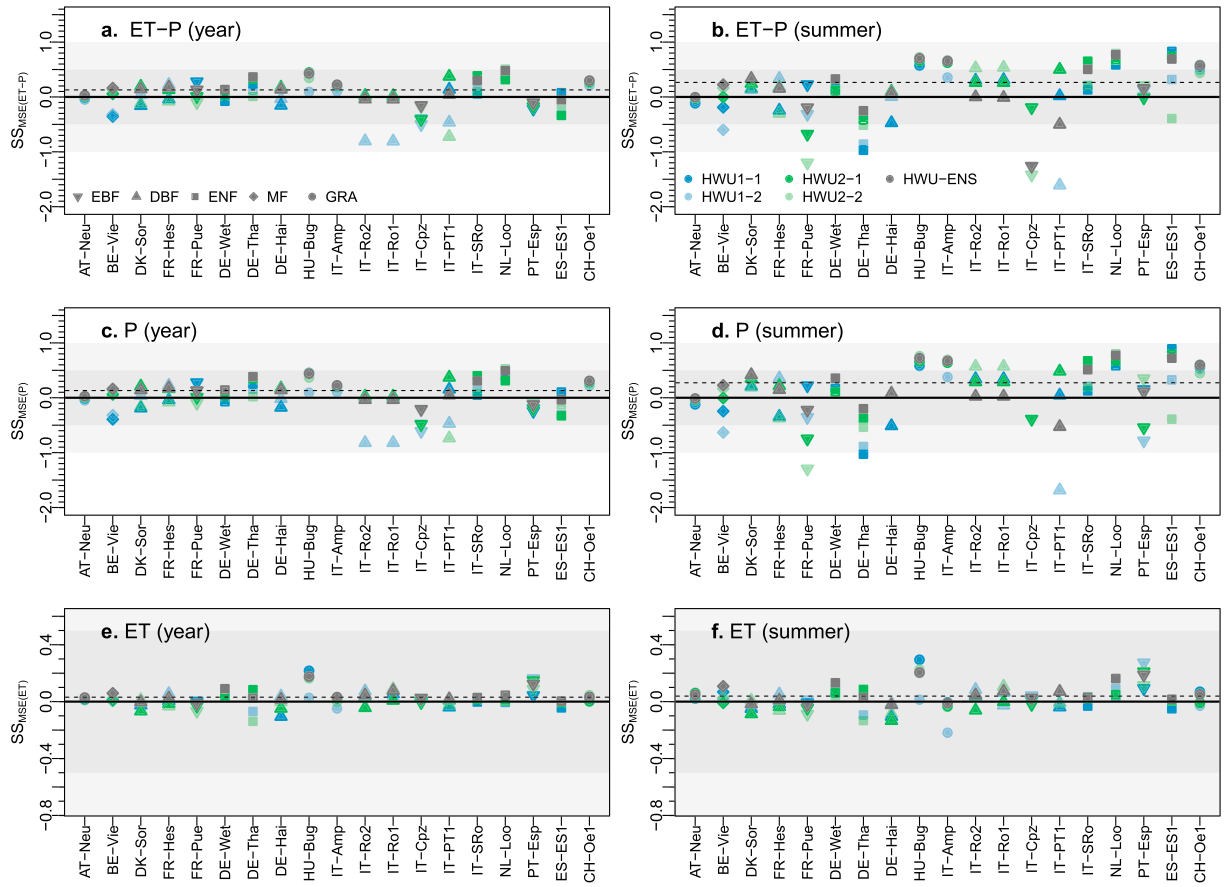


Figure 9. SS_{MSE} for improvements of the water use scenarios over the natural reference run for daily ET-P, ET, and P for the full year and summer 2003. The symbols indicate the dominant land use of the FLUXNET tower. The gray points indicate the average improvements of the water use ensemble mean over the reference run for all stations. The dashed black line shows the average improvement of the water use ensemble over the reference simulation. P = precipitation; ET = evapotranspiration; MSE = mean squared error; EBF = evergreen broadleaf forest; DBF = deciduous broadleaf forest; ENF = evergreen needleleaf forest; MF = mixed forest; GRA = grassland.

~13%, and ET by ~3%. However, single stations can also show considerable deteriorations of skill. Average improvements are larger in summer (ET-P: ~27%, P: ~27%, and ET: ~4%). This improvement can be significantly higher if individual HWU scenarios are inspected. Yet, effects cannot be directly associated with individual plant functional types, regions, or HWU scenarios (cf. Figure S4). The separation of the skill score into ET and P shows that the largest increase in skill is coming from an improved simulation of P, although especially summer skills seem arbitrarily affected due to the chaotic nature of P. Note, while summer ET at PT-Esp is significantly improved through the incorporation of HWU for all water use scenarios (approx. 10% to 40%), the skill of ET-P does not improve, as P skills are not improved (cf. also Figure S5).

As the evaluation at FLUXNET sites indicates major improvements for the simulation of P, individual events are further evaluated with a larger observational data set, to identify an overall added value. Figure 10 shows spatial maps of the SS_{MSE} for daily summer P for HWU-ENS and the individual HWU scenarios. Spatially, the skill improvement varies, and simulated P at single stations is not consistently improved through the incorporation of HWU. Moreover, all stations, including those in less managed regions, indicate changes in skill, which emphasizes the remote impacts of HWU. However, there exist some common patterns of skill improvement for the individual HWU scenarios. The majority of stations over the IB show a negative influence of HWU on the simulation of daily summer P for all HWU scenarios (on average by −9%, cf. Figure S6). In contrast, the majority of the stations over Germany and EA, with a focus on the Eastern parts of the Danube watershed, indicate improved skill (on average by 12.9% over ME). Yet, while the skill of individual HWU scenarios is affected by varying skill through overerestimated and underestimated P and the

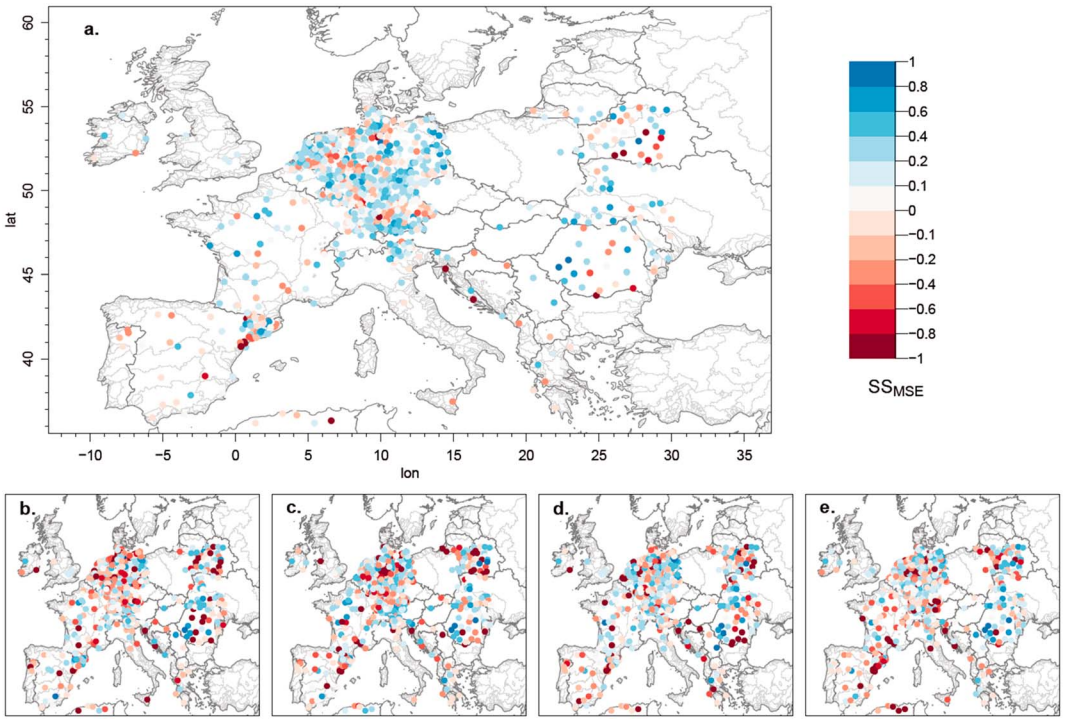


Figure 10. Map of SS_{MSE} for (a) the HWU ensemble mean, (b) HWU1-1, (c) HWU1-2, (d) HWU2-1, and (e) HWU2-2 relative to the NAT simulation for summer 2003. Blue (red) dots indicate an improved (deteriorated) skill in simulating precipitation through the incorporation of HWU. MSE = mean squared error; HWU = human water use.

incorporated uncertainty of HWU (full year: HWU1-1 [−0.5%], HWU1-2 [−0.7%], HWU2-1 [−0.5%], and HWU2-2 [−0.6%]; summer: HWU1-1 [3.1%], HWU1-2 [2.9%], HWU2-1 [6.7%], and HWU2-2 [−2.7%]), the skill improvement of the HWU-ENS indicates an increase of information contained in the simulations through the incorporation of HWU by +7.85% for the full year and 18.96% for summer.

The associated FBI and PSS are very similar between HWU-ENS, the individual HWU scenarios, and NAT (Figures S7 and S8) but indicate that the incorporation of HWU potentially reduces the FBI for VHPD over ME (from 1.92 to 1.61), FR (from 2.16 to 1.84), MD (from 5.84 to 5.25), and EA (from 1.90 to 1.58) but increases over IB (from 8.52 to 9.00). However, overall, large uncertainties remain in attributing local skill improvement, which is potentially related to use of a parameterization scheme for convection and the lack of stations over large parts of Southern Europe.

5. Discussion

5.1. Observational Uncertainty

This study evaluated the uncertainty of bias estimates from simulations using multiple observational data sets, from the local to the watershed scale. The obtained biases of approximately 0.5 mm/day for P and −0.24 mm/day for ET are comparable to other studies (e.g., Katragkou et al., 2015; Kotlarski et al., 2014, 2017; Müller & Seneviratne, 2014; Vautard et al., 2013) and lead to a bias of the net water flux ET-P on the order of −0.29 mm/day. While the incorporation of HWU apparently does not remove these wet biases, this result is subject to observational uncertainty. The influence of observational uncertainty in ET and P on the net water flux bias varies. It has to be highlighted that the capability of the here used observational products to capture irrigation (HWU) effects remains unknown but might at least partly explain their differences. While this uncertainty has previously been identified (e.g., Müller et al., 2013, 2011; Prein & Gobiet, 2017), we have shown here that it exceeds the impact of HWU on the terrestrial water cycle and hence limits the usefulness of these products to validate and potentially calibrate models. The challenges and risks of validating hydrologic- and land surface models with one gridded observational product of, for example, ET are obvious, as in situ measurements remain sparsely located and limited in availability. In

Table 3
Summary of MSE Skill Scores Evaluated in This Study

	Spatial scale	Time scale	SS_{MSE} (P)	SS_{MSE} (ET)	SS_{MSE} (ET-P)
Year	Watershed	Monthly	−1.23%	+0.25%	−1.01%
	Local	Daily	+13.12%	+3.17%	+12.82%
			+7.85%		
Summer	Watershed	Monthly	+11.64%	−0.60%	+8.03%
	Local	Daily	+27.30%	+3.88%	+26.52%
			+18.96%		

Note. The skill scores are illustrated for the HWU ensemble relative to the natural reference simulation. The watershed mean relates to the mean of all observations. The two values for the local, daily scale relate to FLUXNET stations (19 stations; top value) and the ECA&D stations (1,033 stations; bottom value), respectively. Green (red) colored grid cells indicate an average improvement (deterioration) of the MSE through the incorporation of HWU. P = precipitation; ET = evapotranspiration; MSE = mean squared error; HWU = human water use.

general and based on the results of this study, multiple observational data sets should be used and individual simulations should be regarded as single scenarios of a multimodel and multiobservational ensemble, in order to assess uncertainty of water availability under current and future conditions. Note that the true error remains unknown, because the true fluxes are not known everywhere and everytime from direct measurements. Therefore, not only more accurate model simulations are required but also more accurate observational products, which can be used to validate models and allow to assess changes of the terrestrial water cycle exhibiting a smaller uncertainty than the observations.

In addition, the high uncertainty of gridded, observation-based products of ET and P emphasizes the advantage of coupled modeling systems, which do not directly rely on external observational data sets and hence reduce the uncertainty related to, for example, atmospheric forcings (e.g., Biemans et al., 2009). While these models introduce additional biases with respect to, for example, P, which may propagate from one compartment to the other, internal feedback processes and storages are physically consistent. Thus, the origin of biases from observations or simulations cannot be easily identified and constitutes a major challenge for validating coupled modeling systems (Bauer et al., 2015). For future studies, we propose the use of probabilistic approaches to tackle these issues, and the incorporation of observational uncertainty estimates in validation utilizing ensemble approaches with a very large number of ensemble members to simulate the terrestrial water cycle.

5.2. Potential for Added Value of Incorporating HWU

Furthermore, this study assessed the added value of incorporating HWU in the simulation of monthly watershed and daily local fluxes of ET and P, as well as ET-P. The results do not conclusively show that the incorporation of HWU leads to an added value in the simulation of the terrestrial water cycle, as summarized in Table 3. Especially at the watershed scale, the incorporation of HWU may lead to a decreased skill of P and ET-P at the annual time scale and vice versa to a decreased skill of ET in summer. These results do not agree with the daily, local scale analysis, which indicates average improvements for all fluxes for the full year and summer. There are several explanations for this apparent inconsistency. First, the SS_{MSE} for watersheds was calculated using the mean of all observational products, assuming that these represent the true value. Second, there is a lack of in situ stations, which measure ET and P simultaneously. Stations are furthermore not equally and only sparsely distributed across Europe, which may also lead to differences from the watershed to the regional scales.

Furthermore, there are two reasons why the use of HWU-ENS leads to significantly better skill scores than the use of the individual HWU scenarios. As has been mentioned before, the HWU scenarios were not constructed to represent real-world conditions but rather to assess the uncertainty of the associated feedbacks. This might lead to a consistent underestimation and overestimation of individual scenarios, which is averaged out in HWU-ENS. The averaging in HWU-ENS furthermore leads to a much smoother time series with less extremes, which is potentially causing the increase in skill of simulating VHPDs.

The individual HWU scenarios did not improve the bias from the NAT simulation. Here it is noteworthy that differences cannot easily be attributed to local effects of HWU or land use, because we are comparing a 156 km^2 grid box estimate with an in situ measurement over a much smaller footprint based on a nearest neighbor interpolation, irrespective of the land cover.

Given the high uncertainty of the gridded observation-based products and the sparse in situ network, it is hence difficult to conclude that the incorporation of HWU improves the simulation of ET and P. Nevertheless, we find consistent patterns, which indicate the potential of an added value.

6. Summary and Conclusion

This study evaluates the potential added value of integrating HWU in simulations of the continental-scale integrated TerrSysMP on the simulation of the terrestrial water cycle during the heatwave 2003 in Europe.

In the first part, the reference simulation at the seasonal to annual watershed scale is evaluated, considering observational uncertainty. The study finds that observational products of the net moisture flux, ET-P, exhibit a relatively large spatial uncertainty at the watershed scale, which is highest in mountainous watersheds. TerrSysMP simulations reveal a wet bias at the annual and seasonal watershed scale, which is on the order of 0.2 to 2 mm/day but is strongly dependent on the observational product used for validation. The incorporation of HWU does not remove this wet bias in general; however, the observational uncertainty in mountainous watersheds is larger than differences induced by HWU, which precludes the identification of an added value at the watershed scale. While the annual cycle of all fluxes is well reproduced, especially Southern European watersheds exhibit a wet bias, which stems from overestimated P. Watershed averaged ET is especially uncertain during dry and hot summers under predominantly water-limited conditions, during which irrigation is applied. These effects are not consistently captured by the observation-based ET products used in this study, due to the small footprint of the eddy-covariance stations and the time and space limitations of satellite soil moisture to capture irrigation effects (Lawston et al., 2017). Analogously, it remains unclear to which extent eddy covariance observations capture the impact of irrigation. Furthermore, gridded observational P data sets are prone to diverge in regions with a low density of rain gauge observations, which adds to the uncertainty from the different interpolation methods used.

While this study does not conclusively show that the incorporation of HWU leads to an added value in the simulation of the terrestrial water cycle, potential skill improvement for daily ET and P at the local scale is apparent. Unfortunately, colocated measurements of ET and P are scarce, but the 19 FLUXNET stations over Europe indicate an average improvement of daily ET and P values, which is independent from the local land cover. A separation of ET and P shows that the major improvements are in fact arising from an improved simulation of P. A skill assessment with a larger, daily in situ P data set indicates that the incorporation of HWU in an integrated modeling system exhibits a potential added value. In case of TerrSysMP, skill is mainly improved in ME and FR but deteriorated over the IB. In fact, P events over the IB do not show any skill in any of the simulations and show high FBIs for heavy and very heavy P events.

In general, a more accurate simulation of P is needed to improve the simulation of the terrestrial water cycle. The high uncertainty associated with P suggests the largest potential for improvement. This can either be achieved through bias correction (Rojas et al., 2011), an increasing resolution to circumvent the use of simplified convection parameterizations, or data assimilation. The actual added value might still be larger, if assessed at the meso-β scale (Feser, 2006). Moreover, an improved simulation of ET does not necessarily lead to an improved simulation of P, which is highly dependent on the convection parameterization (Hohenegger et al., 2009). In the future, one could bridge the gap between land surface ET and P by using, for example, sounding profiles to assess the propagation of improved land surface states into the atmosphere. Ultimately, remote sensing products should be used to validate long-term changes, also in subsurface storage (using, for example, data from the Gravity Recovery and Climate Experiment), to evaluate an intensification of the terrestrial water cycle due to anthropogenic influences and identify the origin of biases in integrated simulations. Here also longer simulations are needed to assess the long-term impacts of HWU, as the response of the subsurface is much slower.

Acknowledgments
Acknowledgement is made for the use of ECMWF's computing and archive facilities in this research. The authors gratefully acknowledge the computing time granted by the JARA-HPC Vergabegremium and provided on the JARA-HPC Partition part of the supercomputer JUQUEEN at Forschungszentrum Jülich. This research was supported by the SFB/TR32 Patterns in Soil-Vegetation-Atmosphere Systems: Monitoring, Modeling and Data Assimilation project funded by the Deutsche Forschungsgemeinschaft (DFG). Observational data sets are available as indicated in Tables 1 and 2. We thank Brecht Martens for his comments and helpful discussions. The constructive criticism of the Editor Paul Dirmeyer, the associate editor, and of two anonymous reviewers has been greatly beneficial for the quality of the manuscript.

References

Addor, N., Rohrer, M., Furrer, R., & Seibert, J. (2016). Propagation of biases in climate models from the synoptic to the regional scale: Implications for bias adjustment. *Journal of Geophysical Research: Atmospheres*, *121*, 2075–2089. https://doi.org/10.1002/2015JD024040

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., & Reinhardt, T. (2011). Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Monthly Weather Review*, *139*(12), 3887–3905. https://doi.org/10.1175/MWR-D-10-05013.1

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., et al. (2001). FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, *82*(11), 2415–2434. https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Beck, H. E., van Dijk, A. I., Levizzani, V., Schellekens, J., Gonzalez Miralles, D., Martens, B., & de Roo, A. (2017). MSWEP: 3-hourly 0.25 global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589–615. https://doi.org/10.5194/hess-21-589-2017

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., et al. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrology and Earth System Sciences*, *21*(12), 6201–6217. https://doi.org/10.5194/hess-21-6201-2017

Biemans, H., Hutjes, R. W. A., Kabat, P., Strengers, B. J., Gerten, D., & Rost, S. (2009). Effects of precipitation uncertainty on discharge calculations for main river basins. *Journal of Hydrometeorology*, *10*(4), 1011–1025. https://doi.org/10.1175/2008JHM1067.1

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., et al. (2015). Hyper-resolution global hydrological modelling: What is next? "Everywhere and locally relevant". *Hydrological Processes*, *29*(2), 310–320. https://doi.org/10.1002/hyp.10391

Black, E., Blackburn, M., Harrison, G., Hoskins, B., & Methven, J. (2004). Factors contributing to the summer 2003 European heatwave. *Weather*, *59*(8), 217–223. https://doi.org/10.1256/wea.74.04

Bonan, G. B., Lawrence, P. J., Oleson, K. W., Levis, S., Jung, M., Reichstein, M., et al. (2011). Improving canopy processes in the Community Land Model version 4 (CLM4) using global flux fields empirically inferred from FLUXNET data. *Journal of Geophysical Research*, *116*(G2), G02014. https://doi.org/10.1029/2010JG001593

Casanueva, A., Kotlarski, S., Herrera, S., Fernández, J., Gutiérrez, J. M., Boberg, F., et al. (2016). Daily precipitation statistics in a EURO-CORDEX RCM ensemble: Added value of raw and bias-corrected high-resolution simulations. *Climate Dynamics*, *47*(3-4), 719–737. https://doi.org/10.1007/s00382-015-2865-x

Chen, M., Xie, P., Janowiak, J. E., & Arkin, P. A. (2002). Global land precipitation: A 50-yr monthly analysis based on gauge observations. *Journal of Hydrometeorology*, *3*(3), 249–266. https://doi.org/10.1175/1525-7541(2002)003<0249:GLPAYM>2.0.CO;2

Davin, E. L., Maisonnave, E., & Seneviratne, S. I. (2016). Is land surface processes representation a possible weak link in current Regional Climate Models? *Environmental Research Letters*, *11*(7), 074027. https://doi.org/10.1088/1748-9326/11/7/074027

DeAngelis, A., Dominguez, F., Fan, Y., Robock, A., Kustu, M. D., & Robinson, D. (2010). Evidence of enhanced precipitation due to irrigation over the Great Plains of the United States. *Journal of Geophysical Research*, *115*, D15115. https://doi.org/10.1029/2010JD013892

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

de Vrese, P., Hagemann, S., & Claussen, M. (2016). Asian irrigation, African rain: Remote impacts of irrigation. *Geophysical Research Letters*, *43*, 3737–3745. https://doi.org/10.1002/2016GL068146

di Luca, A., de Elía, R., & Laprise, R. (2015). Challenges in the quest for added value of regional climate dynamical downscaling. *Current Climate Change Reports*, *1*(1), 10–21. https://doi.org/10.1007/s40641-015-0003-9

Döll, P., Kaspar, F., & Lehner, B. (2003). A global hydrological model for deriving water availability indicators: Model tuning and validation. *Journal of Hydrology*, *270*(1-2), 105–134. https://doi.org/10.1016/S0022-1694(02)00283-4

Doms, G., & Schättler, U. (2002). A description of the nonhydrostatic regional model LM. Part I: dynamics and numerics. Tech. Rep. Deutscher Wetterdienst, Offenbach am Main, Germany.

FAO. (1988). UNESCO soil map of the world, revised legend. World Resources Report 60, p. 138.

Feser, F. (2006). Enhanced detectability of added value in limited-area model results separated into different spatial scales. *Monthly Weather Review*, *134*(8), 2180–2190. https://doi.org/10.1175/MWR3183.1

Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., et al. (2002). Global land cover mapping from MODIS: Algorithms and early results. *Remote Sensing of Environment*, *83*(1-2), 287–302. https://doi.org/10.1016/S0034-4257(02)00078-0

Gasper, F., Goergen, K., Shrestha, P., Sulis, M., Rihani, J., Geimer, M., & Kollet, S. (2014). Implementation and scaling of the fully coupled Terrestrial Systems Modeling Platform (TerrSysMP v1. 0) in a massively parallel supercomputing environment–a case study on JUQUEEN (IBM Blue Gene/Q). *Geoscientific Model Development*, *7*(5), 2531–2543. https://doi.org/10.5194/gmd-7-2531-2014

Gleeson, T., Marklund, L., Smith, L., & Manning, A. H. (2011). Classifying the water table at regional to continental scales. *Geophysical Research Letters*, *38*, L05401. https://doi.org/10.1029/2010GL046427

Gleeson, T., Smith, L., Moosdorf, N., Hartmann, J., Dürr, H. H., Manning, A. H., et al. (2011). Mapping permeability over the surface of the Earth. *Geophysical Research Letters*, *38*, L02401. https://doi.org/10.1029/2010GL045565

Greve, P., Orlowsky, B., Müller, B., Sheffield, J., Reichstein, M., & Seneviratne, S. I. (2014). Global assessment of trends in wetting and drying over land. *Nature Geoscience*, *7*(10), 716–721. https://doi.org/10.1038/ngeo2247

Guillod, B. P., Orlowsky, B., Miralles, D. G., Teuling, A. J., & Seneviratne, S. I. (2015). Reconciling spatial and temporal soil moisture effects on afternoon rainfall. *Nature Communications*, *6*(1), 6443. https://doi.org/10.1038/ncomms7443

Haddeland, I., Skaugen, T., & Lettenmaier, D. P. (2006). Anthropogenic impacts on continental surface water fluxes. *Geophysical Research Letters*, *33*, L08406. https://doi.org/10.1029/2006GL026047

Harris, I. P. D. J., Jones, P. D., Osborn, T. J., & Lister, D. H. (2014). Updated high-resolution grids of monthly climatic observations—The CRU TS3. 10 Dataset. *International Journal of Climatology*, *34*(3), 623–642. https://doi.org/10.1002/joc.3711

Hazenberg, P., Broxton, P., Gochis, D., Niu, G. Y., Pangle, L. A., Pelletier, J. D., et al. (2016). Testing the hybrid-3-D hillslope hydrological model in a controlled environment. *Water Resources Research*, *52*, 1089–1107. https://doi.org/10.1002/2015WR018106

He, X., Wada, Y., Wanders, N., & Sheffield, J. (2017). Human water management intensifies hydrological drought in California. *Geophysical Research Letters*, *44*, 1777–1785. https://doi.org/10.1002/2016GL071665

Hohenegger, C., Brockhaus, P., Bretherton, C. S., & Schär, C. (2009). The soil moisture–precipitation feedback in simulations with explicit and parameterized convection. *Journal of Climate*, *22*(19), 5003–5020. https://doi.org/10.1175/2009JCLI2604.1

Hohenegger, C., & Schär, C. (2007). Atmospheric predictability at synoptic versus cloud-resolving scales. *Bulletin of the American Meteorological Society*, *88*(11), 1783–1794. https://doi.org/10.1175/BAMS-88-11-1783

Huntington, T. G. (2006). Evidence for intensification of the global water cycle: Review and synthesis. *Journal of Hydrology*, *319*(1-4), 83–95. https://doi.org/10.1016/j.jhydrol.2005.07.003

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Oxford: Wiley.

Jones, J. E., & Woodward, C. S. (2001). Newton–Krylov-multigrid solvers for large-scale, highly heterogeneous, variably saturated flow problems. *Advances in Water Resources*, *24*(7), 763–774. https://doi.org/10.1016/S0309-1708(00)00075-0

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

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., et al. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, *467*(7318), 951–954. https://doi.org/10.1038/nature09396

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., et al. (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research*, *116*, G00J07. https://doi.org/10.1029/2010JG001566

Katragkou, E., García-Díez, M., Vautard, R., Sobolowski, S., Zanis, P., Alexandri, G., et al. (2015). Regional climate hindcast simulations within EURO-CORDEX: evaluation of a WRF multi-physics ensemble. *Geoscientific Model Development*, *8*(3), 603–618. https://doi.org/10.5194/gmd-8-603-2015

Keune, J., Gasper, F., Goergen, K., Hense, A., Shrestha, P., Sulis, M., & Kollet, S. (2016). Studying the influence of groundwater representations on land surface-atmosphere feedbacks during the European heat wave in 2003. *Journal of Geophysical Research: Atmospheres*, *121*, 13,301–13,325. https://doi.org/10.1002/2016JD025426

Keune, J., Sulis, M., Kollet, S., Siebert, S., & Wada, Y. (2018). Human water use impacts on the strength of the continental sink for atmospheric water. *Geophysical Research Letters*, *45*, 4068–4076. https://doi.org/10.1029/2018GL077621

Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, *22*(12), 1441–1453. https://doi.org/10.1002/joc.773

Kollet, S. J., & Maxwell, R. M. (2006). Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model. *Advances in Water Resources*, *29*(7), 945–958. https://doi.org/10.1016/j.advwatres.2005.08.006

Kotlarski, S., Keuler, K., Christensen, O. B., Colette, A., Déqué, M., Gobiet, A., et al. (2014). Regional climate modeling on European scales: A joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geoscientific Model Development*, *7*(4), 1297–1333. https://doi.org/10.5194/gmd-7-1297-2014

Kotlarski, S., Szabó, P., Herrera, S., Räty, O., Keuler, K., Soares, P. M., et al. (2017). Observational uncertainty and regional climate model evaluation: A pan-European perspective. *International Journal of Climatology*. https://doi.org/10.1002/joc.5249

Lawston, P. M., Santanello, J. A., & Kumar, S. V. (2017). Irrigation signals detected from SMAP soil moisture retrievals. *Geophysical Research Letters*, *44*, 11,860–11,867. https://doi.org/10.1002/2017GL075733

Legates, D. R., & Willmott, C. J. (1990). Mean seasonal and spatial variability in gauge-corrected, global precipitation. *International Journal of Climatology*, *10*(2), 111–127. https://doi.org/10.1002/joc.3370100202

Lehner, B., Verdin, K., & Jarvis, A. (2006). *HydroSHEDS technical documentation, version 1.0*. Technical Report (pp. 1–27). Washington, DC: World Wildlife Fund US.

Leuning, R., Zhang, Y. Q., Rajaud, A., Cleugh, H., & Tu, K. (2008). A simple surface conductance model to estimate regional evaporation using MODIS leaf area index and the Penman-Monteith equation. *Water Resources Research*, *44*, W10419. https://doi.org/10.1029/2007WR006562

Liu, X., Tang, Q., Cui, H., Mu, M., Gerten, D., Gosling, S. N., et al. (2017). Multimodel uncertainty changes in simulated river flows induced by human impact parameterizations. *Environmental Research Letters*, *12*(2), 025009. https://doi.org/10.1088/1748-9326/aa5a3a

MacLeod, D. A., Cloke, H. L., Pappenberger, F., & Weisheimer, A. (2016). Improved seasonal prediction of the hot summer of 2003 over Europe through better representation of uncertainty in the land surface. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 79–90. https://doi.org/10.1002/qj.2631

Martens, B., Waegeman, W., Dorigo, W. A., Verhoest, N. E., & Miralles, D. G. (2018). Terrestrial evaporation response to modes of climate variability. *npj Climate and Atmospheric Science*, *1*(1), 43. https://doi.org/10.1038/s41612-018-0053-5

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., et al. (2017). GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*(5), 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017

Maxwell, R. M. (2013). A terrain-following grid transform and preconditioner for parallel, large-scale, integrated hydrologic modeling. *Advances in Water Resources*, *53*, 109–117. https://doi.org/10.1016/j.advwatres.2012.10.001

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., et al. (2017). The future of Earth observation in hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3879–3914. https://doi.org/10.5194/hess-21-3879-2017

McDermid, S. S., Mearns, L. O., & Ruane, A. C. (2017). Representing agriculture in Earth System Models: Approaches and priorities for development. *Journal of Advances in Modeling Earth Systems*, *9*, 2230–2265. https://doi.org/10.1002/2016MS000749

Miralles, D. G., Gentine, P., Seneviratne, S. I., & Teuling, A. J. (2018). Land–atmospheric feedbacks during droughts and heatwaves: State of the science and current challenges. *Annals of the New York Academy of Sciences*, *1436*(1), 19–35. https://doi.org/10.1111/nyas.13912

Miralles, D. G., Holmes, T. R. H., de Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., & Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, *15*(2), 453–469. https://doi.org/10.5194/hess-15-453-2011

Miralles, D. G., Teuling, A. J., van Heerwaarden, C. C., & de Arellano, J. V. G. (2014). Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geoscience*, *7*(5), 345–349. https://doi.org/10.1038/ngeo2141

Miralles, D. G., van den Berg, M. J., Teuling, A. J., & de Jeu, R. A. M. (2012). Soil moisture-temperature coupling: A multiscale observational analysis. *Geophysical Research Letters*, *39*, L21707. https://doi.org/10.1029/2012GL053703

Müller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., et al. (2013). Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis. *Hydrology and Earth System Sciences*, *17*(10), 3707–3720. https://doi.org/10.5194/hess-17-3707-2013

Müller, B., & Seneviratne, S. I. (2014). Systematic land climate and evapotranspiration biases in CMIP5 simulations. *Geophysical Research Letters*, *41*, 128–134. https://doi.org/10.1002/2013GL058055

Müller, B., Seneviratne, S. I., Jimenez, C., Corti, T., Hirschi, M., Balsamo, G., et al. (2011). Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations. *Geophysical Research Letters*, *38*, L06402. https://doi.org/10.1029/2010GL046230

Müller-Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., & Döll, P. (2014). Sensitivity of simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration. *Hydrology and Earth System Sciences*, *18*(9), 3511–3538. https://doi.org/10.5194/hess-18-3511-2014

Müller-Schmied, H., Adam, L., Eisner, S., Fink, G., Flörke, M., Kim, H., et al. (2016). Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use. *Hydrology and Earth System Sciences*, *20*(7), 2877–2898. https://doi.org/10.5194/hess-20-2877-2016

Nazemi, A., & Wheater, H. S. (2015). On inclusion of water resource management in Earth system models—Part 1: Problem definition and representation of water demand. *Hydrology and Earth System Sciences*, *19*(1), 33–61. https://doi.org/10.5194/hess-19-33-2015

Oki, T., & Kanae, S. (2006). Global hydrological cycles and world water resources. *Science*, *313*(5790), 1068–1072. https://doi.org/10.1126/science.1128845

Oleson, K. W., Niu, G. Y., Yang, Z. L., Lawrence, D. M., Thornton, P. E., Lawrence, P. J., et al. (2008). Improvements to the Community Land Model and their impact on the hydrological cycle. *Journal of Geophysical Research*, *113*, G01021. https://doi.org/10.1029/2007JG000563

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

Orth, R., Dutra, E., & Pappenberger, F. (2016). Improving weather predictability by including land surface model parameter uncertainty. *Monthly Weather Review*, *144*(4), 1551–1569. https://doi.org/10.1175/MWR-D-15-0283.1

Pokhrel, Y. N., Koirala, S., Yeh, P. J. F., Hanasaki, N., Longuevergne, L., Kanae, S., & Oki, T. (2015). Incorporation of groundwater pumping in a global land surface model with the representation of human impacts. *Water Resources Research*, *51*, 78–96. https://doi.org/10.1002/2014WR015602

Prein, A. F., & Gobiet, A. (2017). Impacts of uncertainties in European gridded precipitation observations on regional climate analysis. *International Journal of Climatology*, *37*(1), 305–327. https://doi.org/10.1002/joc.4706

Prein, A. F., Gobiet, A., Suklitsch, M., Truhetz, H., Awan, N. K., Keuler, K., & Georgievski, G. (2013). Added value of convection permitting seasonal simulations. *Climate Dynamics*, *41*(9-10), 2655–2677. https://doi.org/10.1007/s00382-013-1744-6

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., et al. (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of Geophysics*, *53*, 323–361. https://doi.org/10.1002/2014RG000475

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., et al. (2016). Multiscale and multivariate evaluation of water fluxes and states over European river basins. *Journal of Hydrometeorology*, *17*(1), 287–307. https://doi.org/10.1175/JHM-D-15-0054.1

Ritter, B., & Geleyn, J. F. (1992). A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. *Monthly Weather Review*, *120*(2), 303–325. https://doi.org/10.1175/1520-0493(1992)120<0303:ACRSFN>2.0.CO;2

Rojas, R., Feyen, L., Dosio, A., & Bavera, D. (2011). Improving pan-European hydrological simulation of extreme events through statistical bias correction of RCM-driven climate simulations. *Hydrology and Earth System Sciences*, *15*(8), 2599–2620. https://doi.org/10.5194/hess-15-2599-2011

Rudolf, B., Beck, C., Grieser, J., & Schneider, U. (2005). Global precipitation analysis products of the GPCC. Global Precipitation Climatology Centre, Deutscher Wetterdienst, Offenbach am Main, Germany.

Rummukainen, M. (2016). Added value in regional climate modeling. *Wiley Interdisciplinary Reviews: Climate Change*, *7*(1), 145–159. https://doi.org/10.1002/wcc.378

Schär, C., & Jendritzky, G. (2004). Climate change: Hot news from summer 2003. *Nature*, *432*(7017), 559–560. https://doi.org/10.1038/432559a

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., & Ziese, M. (2011). GPCC full data reanalysis version 6.0 at 0.5: Monthly land-surface precipitation from rain-gauges built on GTS-based and historic data. Global Precipitation Climatology Centre, Deutscher Wetterdienst, Offenbach am Main, Germany. https://doi.org/10.5676/DWD_GPCC/FD_M_V7_050.

Shrestha, P., Sulis, M., Masbou, M., Kollet, S., & Simmer, C. (2014). A scale-consistent terrestrial systems modeling platform based on COSMO, CLM, and ParFlow. *Monthly Weather Review*, *142*(9), 3466–3483. https://doi.org/10.1175/MWR-D-14-00029.1

Siebert, S., Burke, J., Faures, J. M., Frenken, K., Hoogeveen, J., Döll, P., & Portmann, F. T. (2010). Groundwater use for irrigation—A global inventory. *Hydrology and Earth System Sciences*, *14*(10), 1863–1880. https://doi.org/10.5194/hess-14-1863-2010

Siebert, S., & Döll, P. (2010). Quantifying blue and green virtual water contents in global crop production as well as potential production losses without irrigation. *Journal of Hydrology*, *384*(3-4), 198–217. https://doi.org/10.1016/j.jhydrol.2009.07.031

Sörensson, A. A., & Ruscica, R. C. (2018). Intercomparison and uncertainty assessment of nine evapotranspiration estimates over South America. *Water Resources Research*, *54*, 2891–2908. https://doi.org/10.1002/2017WR021682

Sulis, M., Keune, J., Shrestha, P., Simmer, C., & Kollet, S. J. (2018). Quantifying the impact of subsurface-land surface physical processes on the predictive skill of subseasonal mesoscale atmospheric simulations. *Journal of Geophysical Research: Atmospheres*, *123*, 9131–9151. https://doi.org/10.1029/2017JD028187

Sulis, M., Williams, J. L., Shrestha, P., Diederich, M., Simmer, C., Kollet, S. J., & Maxwell, R. M. (2017). Coupling groundwater, vegetation, and atmospheric processes: A comparison of two integrated models. *Journal of Hydrometeorology*, *18*(5), 1489–1511. https://doi.org/10.1175/JHM-D-16-0159.1

Swenson, S. C., & Lawrence, D. M. (2014). Assessing a dry surface layer-based soil resistance parameterization for the Community Land Model using GRACE and FLUXNET-MTE data. *Journal of Geophysical Research: Atmospheres*, *119*, 10,299–10,312. https://doi.org/10.1002/2014JD022314

Thiery, W., Davin, E. L., Lawrence, D. M., Hirsch, A. L., Hauser, M., & Seneviratne, S. I. (2017). Present-day irrigation mitigates heat extremes. *Journal of Geophysical Research: Atmospheres*, *122*, 1403–1422. https://doi.org/10.1002/2016JD025740

Tiedtke, M. I. C. H. A. E. L. (1989). A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, *117*(8), 1779–1800. https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2

Trambauer, P., Dutra, E., Maskey, S., Werner, M., Pappenberger, F., van Beek, L. P. H., & Uhlenbrook, S. (2014). Comparison of different evaporation estimates over the African continent. *Hydrology and Earth System Sciences*, *18*(1), 193–212. https://doi.org/10.5194/hess-18-193-2014

Tuinenburg, O. A., & de Vries, J. P. R. (2017). Irrigation patterns resemble ERA-Interim Reanalysis soil moisture additions. *Geophysical Research Letters*, *44*, 10,341–10,348. https://doi.org/10.1002/2017GL074884

Valcke, S. (2013). The OASIS3 coupler: A European climate modelling community software. *Geoscientific Model Development*, *6*(2), 373–388 . https://doi.org/10.5194/gmd-6-373-2013

Vautard, R., Gobiet, A., Jacob, D., Belda, M., Colette, A., Déqué, M., et al. (2013). The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project. *Climate Dynamics*, *41*(9-10), 2555–2575. https://doi.org/10.1007/s00382-013-1714-z

Veldkamp, T. I. E., Zhao, F., Ward, P. J., de Moel, H., Aerts, J. C. J. H., Schmied, H. M., et al. (2018). Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: A multi-model validation study. *Environmental Research Letters*, *13*(5), 055008. https://doi.org/10.1088/1748-9326/aab96f

Voisin, N., Li, H., Ward, D., Huang, M., Wigmosta, M., & Leung, L. R. (2013). On an improved sub-regional water resources management representation for integration into Earth system models. *Hydrology and Earth System Sciences*, *17*(9), 3605–3622. https://doi.org/10.5194/hess-17-3605-2013

von Storch, H., Langenberg, H., & Feser, F. (2000). A spectral nudging technique for dynamical downscaling purposes. *Monthly Weather Review*, *128*(10), 3664–3673. https://doi.org/10.1175/1520-0493(2000)128<3664:ASNTFD>2.0.CO;2

Vörösmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global water resources: Vulnerability from climate change and population growth. *Science*, *289*(5477), 284–288. https://doi.org/10.1126/science.289.5477.284

Wada, Y., de Graaf, I. E., & van Beek, L. P. (2016). High-resolution modeling of human and climate impacts on global water resources. *Journal of Advances in Modeling Earth Systems*, *8*, 735–763. https://doi.org/10.1002/2015MS000618

Wada, Y., van Beek, L. P. H., & Bierkens, M. F. (2012). Nonsustainable groundwater sustaining irrigation: A global assessment. *Water Resources Research*, *48*, W00L06. https://doi.org/10.1029/2011WR010562

Wada, Y., Bierkens, M. F. P., de Roo, A., Dirmeyer, P. A., Famiglietti, J. S., Hanasaki, N., et al. (2017). Human–water interface in hydrological modelling: Current status and future directions. *Hydrology and Earth System Sciences*, *21*(8), 4169–4193. https://doi.org/10.5194/hess-21-4169-2017

Wagener, T., Sivapalan, M., Troch, P. A., McGlynn, B. L., Harman, C. J., Gupta, H. V., et al. (2010). The future of hydrology: An evolving science for a changing world. *Water Resources Research*, *46*, W05301. https://doi.org/10.1029/2009WR008906

Wanders, N., & Wada, Y. (2015). Human and climate impacts on the 21st century hydrological drought. *Journal of Hydrology*, *526*, 208–220. https://doi.org/10.1016/j.jhydrol.2014.10.047

Wang, A., Zeng, X., & Guo, D. (2016). Estimates of global surface hydrology and heat fluxes from the Community Land Model (CLM4. 5) with four atmospheric forcing datasets. *Journal of Hydrometeorology*, *17*(9), 2493–2510. https://doi.org/10.1175/JHM-D-16-0041.1

Weisheimer, A., Doblas-Reyes, F. J., Jung, T., & Palmer, T. N. (2011). On the predictability of the extreme summer 2003 over Europe. *Geophysical Research Letters*, *38*, L05704. https://doi.org/10.1029/2010GL046455

Willmott, C. J., & Robeson, S. M. (1995). Climatologically aided interpolation (CAI) of terrestrial air temperature. *International Journal of Climatology*, *15*(2), 221–229. https://doi.org/10.1002/joc.3370150207

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., et al. (2002). Energy balance closure at FLUXNET sites. *Agricultural and Forest Meteorology*, *113*(1-4), 223–243. https://doi.org/10.1016/S0168-1923(02)00109-0

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

Zhang, Y., Peña-Arancibia, J. L., McVicar, T. R., Chiew, F. H. S., Vaze, J., Liu, C., et al. (2016). Multi-decadal trends in global terrestrial evapotranspiration and its components. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep19124