Spatial Sensitivity Analysis of Simulated Land Surface Patterns in a Catchment Model Using a Set of Innovative Spatial Performance Metrics

JULIAN KOCH

Department of Hydrology, Geological Survey of Denmark and Greenland, and Department of Geosciences and Natural Resources Management, University of Copenhagen, Copenhagen, Denmark

GORKA MENDIGUREN

Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark

GREGOIRE MARIETHOZ

Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland

SIMON STISEN

Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark

ABSTRACT

Distributed hydrological models simulate states and fluxes of water and energy in the terrestrial hydrosphere at each cell. The predicted spatial patterns result from complex nonlinear relationships and feedbacks. Spatial patterns are often neglected during the modeling process, and therefore a spatial sensitivity analysis framework that highlights their importance is proposed. This study features a comprehensive analysis of spatial patterns of actual evapotranspiration (ET) and land surface temperature (LST), with the aim of quantifying the extent to which forcing data and model parameters drive these patterns. This framework is applied on a distributed model [MIKE Système Hydrologique Européen (MIKE SHE)] coupled to a land surface model [Shuttleworth and Wallace–Evapotranspiration (SW-ET)] of a catchment in Denmark. Twenty-two scenarios are defined, each having a simplified representation of a potential driver of spatial variability. A baseline model that incorporates full spatial detail is used to assess sensitivity. High sensitivity can be attested in scenarios where the simulated spatial patterns differ significantly from the baseline. The core novelty of this study is that the analysis is based on a set of innovative spatial performance metrics that enable a reliable spatial pattern comparison. Overall, LST is very sensitive to air temperature and wind speed whereas ET is rather driven by vegetation. Both are sensitive to groundwater coupling and precipitation. The conclusions may be limited to the selected catchment and to the applied modeling system, but the suggested framework is generically relevant for the modeling community. While the applied metrics focus on specific spatial information, they partly exhibit redundant information. Thus, a combination of metrics is the ideal approach to evaluate spatial patterns in models outputs.

1. Introduction

Hydrological states and fluxes at the land surface govern nearly all processes at the land–atmosphere interface and are driven by a complex interplay of climate, vegetation, and surface and subsurface dynamics. Understanding the spatiotemporal variability of the terrestrial hydrological cycle is hence a grand scientific and societal challenge addressing issues of, for example, climate change, drought and flood risk, or water resources management. In this regard, spatially distributed, process-based modeling is a key scientific tool to predict hydrological variability at various scales (Fang et al. 2015; Sheffield et al. 2014; Xia et al. 2016). Current modeling frameworks are modular and hydrological processes are either coupled by means of boundary layers or fully integrated, which is an essential feature to adequately represent the hydrological system with its

---

Corresponding author e-mail: Julian Koch, juko@geus.dk

dynamic feedbacks and complex nonlinear relationships (Clark et al. 2015; Samaniego et al. 2010; Shrestha et al. 2014). For example, Williams and Maxwell (2011) found that the spatial variability of subsurface parameters propagates through the system via soil moisture into the atmosphere by affecting evapotranspiration (ET) patterns at the land–atmosphere interface. Furthermore, Butts et al. (2014) and Larsen et al. (2016b) stressed the importance of coupling a groundwater model to a climate model via land surface processes by identifying diverging precipitation patterns between the climate model alone and the coupled model. The importance of considering subsurface processes when modeling land surface variability is also stressed by Tian et al. (2012), who identified distinct evapotranspiration sensitivity to groundwater depth and hydraulic conductivity parameters in a fully coupled model.

Hydrological land surface variables have distinct spatial variability, and different spatial patterns emerge at different time scales. This poses rigorous requirements to the quality of the forcing data and accuracy of the model parameters, as indicated by Bierkens et al. (2015), Wood et al. (2011), and others. Distributed process–based hydrological models are typically highly parameterized and most parameters have a spatial dimension. Many studies draw information from remote sensing data to infer meaningful spatial parameter fields as well as forcing data (Mascaro et al. 2015; Mitchell et al. 2004; Stisen et al. 2011). More frequently, remote sensing data are incorporated to spatially validate or calibrate distributed models (Corbari and Marco 2014; Koch et al. 2015; Wanders et al. 2014). Despite these previous efforts, there is limited knowledge regarding the extent to which spatial model inputs drive the simulated spatial patterns at the land surface. Chaney et al. (2015), Qiu et al. (2014), and others studied spatial variability of soil moisture and its underlying processes from a modeling perspective, but the remaining land surface variables are typically overlooked. It is essential to comprehend the importance of potential drivers of spatial variability in order to better predict the holistic spatial variability of all relevant hydrological states and their fluxes at the land–atmosphere interface. Understanding the most critical processes will equip the modeling community with guidance to better diagnose spatial model deficiencies and to efficiently incorporate spatial detail (e.g., by means of remote sensing) to model parameters and forcing.

This study aims at making a contribution toward this demand by conducting a comprehensive scenario-based spatial sensitivity analysis of a distributed, process-based catchment model in Denmark. Traditionally, the term sensitivity refers to the partial derivatives of the model's response around a local point in the parameter space (Razavi and Gupta 2015). In our study, a baseline model with high spatial detail in all its input parameters and forcing builds the reference for the scenario analysis. In the following, each scenario contains deteriorated spatial detail of a potential driver of spatial variability, and the effect on the simulated patterns is investigated. The focus lies on ET and land surface temperature (LST), and the pattern response is evaluated for each scenario with respect to the baseline model. Therefore, the term sensitivity should not be understood in the traditional way, because partial derivatives cannot be calculated as such. A scenario is rated as sensitive if it significantly alters the spatial patterns of LST and ET with respect to the baseline model. The way this scenario analysis is set up allows us to draw conclusions of spatial sensitivities in the modeling system, and results may not be directly imposed on the natural system or on other catchments. However, the suggested framework can be regarded as a generic contribution to the spatially distributed modeling community, because it is directly transferable to other catchments and modeling systems.

Performing a sensitivity analysis calls for defining a metric to determine whether two models are spatially different (indicating high sensitivity to the underlying parameters) or spatially similar (low sensitivity). To provide a reliable measure of spatial similarity, it is necessary to include spatial performance metrics that allow a true pattern comparison that goes beyond simple cell-to-cell comparisons (Wealands et al. 2005). The main novelty in this study is the application of a set of innovative metrics that includes 1) a multiple-point compatibility (MPC) algorithm (Pérez et al. 2014), 2) a connectivity (CON) analysis (Renard and Allard 2013), 3) an empirical orthogonal function (EOF) analysis (Koch et al. 2015), 4) a fractions skill score (FSS) assessment (Roberts and Lean 2008), and 5) a variogram analysis (Chiles and Delfiner 2009). All of the given metrics are bias insensitive, which we regard as a key feature when assessing spatial patterns. The intention is not to deny the importance of the bias as a spatial performance metric, because a systematic bias, which may leave the spatial pattern unchanged, still provides valuable insights. Ideally, the bias, because it is so easy to compute and interpret, should be consulted as an auxiliary metric next to true spatial pattern metrics (Koch et al. 2015). The spatial metrics quantify the spatial similarity of a scenario with respect to the baseline model, and the similarity scores are utilized as an indicator for spatial sensitivity. In general, each metric focuses on specific spatial information, and hence a pattern evaluation can be expected to result in diverging results. We investigate this by means of a linear
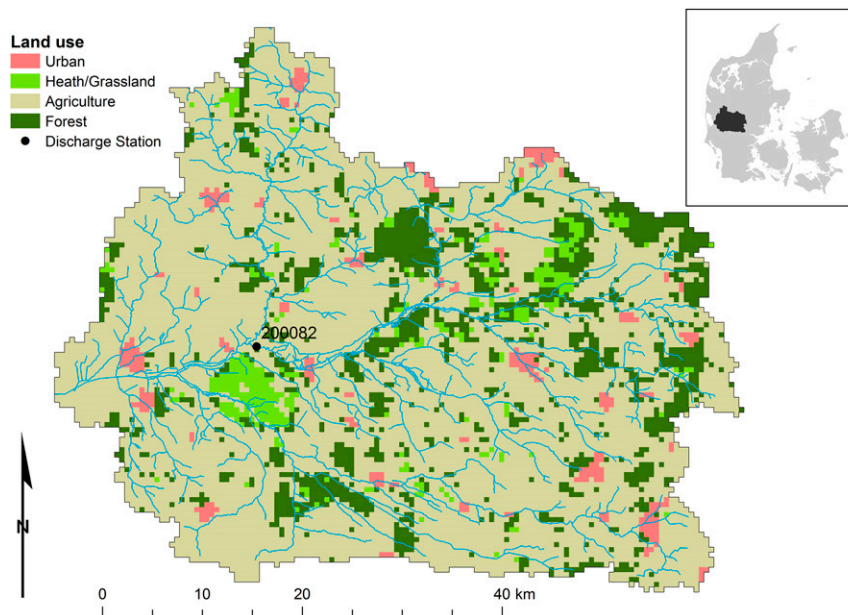
FIG. 1. The land-use map of the Skjern River catchment at 500 m resolution, including one discharge station.

regression analysis with the aim to quantify uniqueness and redundancy in the information content of the metrics. The set of applied metrics is a timely contribution to the modeling community where the spatial-pattern-oriented model evaluation using remote sensing data receives growing attention.

The overall aims of this study are 1) to introduce a set of innovative bias-insensitive, spatial performance metrics that allow a meaningful comparison of spatial patterns; 2) to apply these metrics in a scenario-based spatial sensitivity analysis of simulated patterns of ET and LST of a distributed, process-based catchment model; and 3) to investigate uniqueness and redundancy in the spatial information content of the respective metrics.

2. Study site and data

a. Skjern catchment

The Skjern River catchment is located in the western part of the Danish peninsula and covers 2500 km². The catchment serves as the main research site of the Danish hydrological observatory (HOBE) project and has been studied intensively for almost a decade (Jensen and Illangasekare 2011). With a maritime climate, mean annual precipitation amounts to 990 mm and mean reference evapotranspiration accounts for 575 mm of the annual water balance. Soils are predominately sandy and originate from glacial outwash plains with

intertwined sections of clay and till. The topography slopes gently from the coast on the western side to 125 m MSL at the eastern side of the catchment. Figure 1 depicts the land-use map at the Skjern catchment, where agriculture is the dominant class followed by forest and heath. The map does not indicate the extensive monitoring network, which consists of state-of-the-art instrumentations for continuous measurements of eddy covariance towers (Ringgaard et al. 2011), a wireless soil moisture network (Bircher et al. 2012), and all other relevant hydrological states and fluxes throughout the catchment.

b. Hydrological model

The hydrological model of the Skjern River catchment is set up in the MIKE Système Hydrologique Européen (MIKE SHE) modeling system (Abbott et al. 1986) and comprises fully coupled modules of 3D saturated flow (finite difference), 1D unsaturated flow (Richards' equation), river routing, and overland flow. The traditional MIKE SHE setup is extended by an additional coupling of a land surface model (SW-ET) that solves the diurnal energy balance (Shuttleworth and Wallace 1985). The land surface model provides a more physical description of the processes at the land surface–atmosphere interface than the more conceptual model based on the potential evapotranspiration approach (Kristensen and Jensen 1975). This extension requires hourly climate forcing and time steps in order to reflect the complex diurnal variability correctly (Overgaard 2005).

TABLE 1. Overview of the calibration results. The calibration scheme incorporates five observational datasets and 11 objective functions. The objective functions are defined as RMSE, slope of the linear regression line, bias, and correlation coefficient. The values are based on averages; for example, the RMSE value for discharge represents the average of all eight stations. Fields marked with an em dash are not considered as objective functions in the calibration.

Variable	No. of obs	RMSE	Slope	Bias	R	Extra
Discharge ($m^3\,s^{-1}$)	11 680	0.95	—	−0.24	—	8 stations
Hydraulic head (m)	618	3.47	—	−0.36	—	11 geological layers
Soil moisture (%)	4724	4.90	0.64	—	—	30 stations
ET (mm)	333	0.66	—	−0.07	—	2 flux towers
LST (K)	28 575	2.64	—	0.08	0.33	28 days with coverage >95%

In addition, a vegetation parameterization scheme is derived from the Moderate Resolution Imaging Spectroradiometer (MODIS), and details are presented in the following section. The Skjern model is set up for a period from 2002 to 2010 at 500 m resolution, and the spatial sensitivity analysis focuses exclusively on the last 3 years. At 500 m resolution, relevant small-scale processes such as hillslope–riparian–stream interactions are not reflected accordingly. However, reliable parameters and forcing datasets as well as remote sensing observations are not available at finer scales, which limit the model's predictability below 500 m resolution.

The model is calibrated using a multiconstraint framework that is briefly introduced in Table 1. The calibration is based on five independent observational datasets: discharge, hydraulic head, actual evapotranspiration, soil moisture, and satellite-derived LST. In total, 11 model parameters are included in the optimization. They are related to the land surface, saturated zone, and unsaturated zone modules. The parameter values are estimated using a global optimization scheme based on covariance matrix adaptation estimation strategies (CMA-ES) as introduced by Hansen and Ostermeier (2001) and later applied by Bayer and Finkel (2007) to calibrate a groundwater model against concentration. The Parameter Estimation (PEST; Doherty 2016) has a built-in CMA-ES module, which is used for this calibration task. The calibration reduces errors in 11 objective functions, and the final values are stated in Table 1. In general, the root-mean-square error (RMSE) and bias are selected as metrics for the five different observational variables. However, a few exceptions are made for soil moisture and LST where the slope of the linear regression line or the correlation coefficient are considered as well. Current calibration frameworks of catchment models target the fitting of time series, and hence objective functions of discharge, hydraulic head, and soil moisture can effectively be reduced. Although remotely sensed spatial patterns of LST are included in the calibration, the spatial performance shows limited improvements, because the

calibration design is more efficient at generating temporal, as opposed to spatial, coherence. Further details of the calibration are not covered by this study because the focus lies on the successive spatial sensitivity analysis. Hence, the purpose of the calibration is to grant a baseline scenario that is consistent with observations as shown by Table 1.

c. Remote sensing data

The MODIS sensor flies on board the polar-orbiting platforms *Terra* and *Aqua* and has four overpasses per day at midlatitudes. For the vegetation parameterization of the distributed Skjern model, we incorporated different MODIS products obtained from the NASA Land Processes Distributed Active Archive Center (LP DAAC). Table 2 presents a brief description of the remote sensing products used in this study. During some periods, low sun zenith angles and snow cover cause low coverage of the MODIS leaf area index (LAI) product (MCD15A2) data marked with best-quality flags. Therefore, we build a site-specific empirical relationship between the LAI and the normalized difference vegetation index (NDVI; Rouse et al. 1973). The NDVI is an index sensitive to biomass and chlorophyll content of the vegetation and has been used extensively to monitor different properties of the vegetation by establishing empirical relationships between the index and the variable of interest. For example, Boegh et al. (2004) derived LAI data for Denmark from NDVI using remote sensing data with an exponential function and in situ LAI data for Denmark. The Nadir Bidirectional Reflectance Distribution Function (BRDF)-Adjusted Reflectance (NBAR) from the MCD43B4 product is used to calculate the NDVI by

$$\mathrm{NDVI} = \frac{B_2 - B_1}{B_2 + B_1}, \tag{1}$$

where B_x indicates the MODIS band number. In the following, TIMESAT, version 3.2 (Jönsson and Eklundh 2004, 2002), is utilized to smooth the dataset and to

TABLE 2. Description of the applied MODIS products.

Product	Description	Spatial resolution	Temporal resolution
MCD15A2	LAI and photosynthetically active radiation	1 km	Every 8 days with 16 days of acquisition
MCD43B4	NBAR	1 km	Every 8 days with 16 days of acquisition
MOD09A1	Daily surface reflectance	500 m	Every 8 days

ensure consistency in the data. TIMESAT fits a function to the data based on least squares criteria. The Savitzky–Golay filter is selected as it preserves the maximum and minimum values of the original dataset (Savitzky and Golay 1964).

The empirical relationship to obtain the fitting equation is based on the input LAI data used in the Danish National Water Resources model (Stisen et al. 2012). In the Skjern catchment, the main land use is agriculture, and thus the fitting equation is based on mean LAI and NDVI values at collocated agricultural cells. For this study we applied a similar approach as in Boegh et al. (2004) by adjusting the coefficients to match data:

$$\text{LAI} = 0.0633 \cdot e^{5.524(\text{NDVI})}. \tag{2}$$

Spatially distributed root depth (RD) is derived using a similar approach to Andersen et al. (2002), where RD is calculated by

$$\text{RD}_i = \text{RD}_{max} \frac{\text{LAI}_i}{\text{LAI}_{max}}, \tag{3}$$

where RD_i is RD in pixel i, LAI_i is the LAI in that cell, and LAI_{max} and RD_{max} indicate the maximum values at i. This equation is modified and LAI is substituted by NDVI,

$$\text{RD}_i = \text{RD}_{max} \frac{\text{NDVI}_i}{\text{NDVI}_{max}}, \tag{4}$$

and used for forested land-cover types, poorly vegetated areas, and urban areas. It should be noted that RD is an effective parameter with limited physical relevancy. For instance, Eq. (4) equips forest cells with a temporal-varying RD that is contrary to our physical understanding, but it compensates for mixed land-use cells and scaling issues. RD of croplands and grasslands, denoted as $\text{RD}_{(agri)}$, is estimated by implementing some modifications to Eq. (4). The RD_{max} varies depending on the soil type, and therefore an empirical relation between the plant available water (PAW) and RD_{max} is established for the different soil types. PAW reflects the difference in soil water content between the field capacity and wilting point. This equation is then included as a substitute of RD_{max}. To allow RD to reach zero, the second term in Eq. (4) is normalized by including NDVI_{min} in the equation:

$$\text{RD}_{(agri)i} = 0.12 \cdot e^{[10(\text{PAW}_i)]} \frac{\text{NDVI}_i - \text{NDVI}_{min}}{\text{NDVI}_{max} - \text{NDVI}_{min}}. \tag{5}$$

Following Stisen et al. (2011), the vegetation height (VH) is calculated by assuming a simple linear regression,

$$\text{VH} = 0.14(\text{LAI}) + 0.09, \tag{6}$$

and forest cells are set to a year-round constant VH of 7 m. The shortwave broadband albedo is estimated every 8 days at 1-km spatial resolution from the individual bands of 8-day surface reflectance product (MOD09A1) containing atmospherically corrected narrowband MODIS reflectance, using the algorithm by Liang et al. (2003).

d. Model scenarios

The highly parameterized Skjern model is based on a distinct degree of spatial detail, and Table 3 lists all parameters and forcing data that have a spatial dimension. In total, 43 rain gauges within and around the catchment are utilized to generate daily interpolated precipitation fields. The remaining climate variables are available from up to 16 climate stations with temporal resolutions of 1 and 3 h. The soil hydraulic parameters for the van Genuchten model, which describe the soil water retention curve of unsaturated soils, are given for 16 soil types in the catchment, and the spatial distribution is derived from a national soil map at 500 m resolution. The geological model of the top 3 m consists of three units (sand, clay, and till) that have individual values for hydraulic conductivity, specific storage, and specific yield. Deeper layers of the geological model are not addressed in this study because of their limited coupling to surface processes. LAI, root depth, albedo, and vegetation height are all derived from remote sensing data as described previously. The root distribution is derived from national bulk density maps of the top soil layer. Values for the remaining vegetation parameters (interception, stomata resistance, leaf width, and light extinction) are given for the three major classes (agriculture, forest, and grassland/heath) of the land-use map (Fig. 1) following literature values found in Larsen et al. (2016a).

The baseline model considers the full integration of all spatially distributed parameters and forcing data

TABLE 3. Overview of the 22 scenarios and the model forcings and parameters that have a spatial dimension. Further, the scenarios are grouped with respect to the model module they affect.

Module	Scenario No.	Forcing data/parameter	Original resolution	Spatial perturbation	Temporal
Climate	1	Air temperature	2 km	Set constant	Yes
	2	Air pressure			
	3	Wind speed			
	4	Global radiation			
	5	Humidity			
	6	Precipitation	500 m		
Soil and geology	7	Soil types	500 m	Set to predominant class	No
	8	Top 3-m geology			
	9	Disable groundwater coupling		Fixed head below RD	
Vegetation	10	LAI	1 km	10-km smoothing	Yes
	11	RD			
	12	Albedo			
	13	Vegetation height			
	14	LAI		Set constant	
	15	RD			
	16	Albedo			
	17	Vegetation height			
	18	Root distribution (AROOT)	500 m		No
	19	Interception (Cint)			
	20	Min stomata resistance (RSC)			
	21	Leaf width (LW)			
	22	Light extinction (Kext)			

mentioned above. Table 3 describes the 22 scenarios that are applied in this study, each one including a deterioration of spatial detail of the baseline model by perturbing one potential driver at a time. The premise is that if a perturbation in spatial input alters the simulated spatial patterns of LST and ET significantly, the model is deemed sensitive to that specific driver. Vice versa, little change in the simulated patterns of a scenario indicates limited sensitivity. The MODIS-derived vegetation parameters have the highest degree of spatial heterogeneity; therefore, the spatial variability is reduced in two steps: 1) by a moving average 10-km smoothing filter and 2) by spatially averaging to a constant value for the entire catchment. Scenario 9, which addresses the sensitivity of groundwater coupling, represents a special case. The spatial variability of potential drivers remains unchanged, but the conceptual model setup is simplified by disabling the saturated zone and setting a constant groundwater head boundary below the root depth to decouple the land surface from potential groundwater influences.

Interactions and feedbacks between potential drivers are disregarded because the applied sensitivity analysis is based on one-at-a-time perturbations. The underlying assumption that there exist no interactions between potential drivers and the fact that the changes in model response are always related to the same baseline may limit the conclusion in terms of identifying global

sensitivities. However, we are not aware of any formal approach of how to assess the global sensitivity of spatial patterns of input parameters; instead, it is typically applied for specific parameter values (Razavi and Gupta 2015; Sobol' and Kucherenko 2009). In this way, this study treats the term sensitivity as practical rather than mathematically sound, with the aim to guide model builders in setting up spatially distributed models.

The baseline model and the scenarios perform reasonably well with Nash–Sutcliffe efficiency values ranging from 0.81 to 0.87 for station 200082 (Fig. 1). This underlines that all scenarios are capable of simulating the temporal discharge dynamics adequately while maintaining a correct water balance as well. Furthermore, this stresses the limited sensitivity of discharge to the spatial patterns of catchment inherent processes (Koch et al. 2016a; Pokhrel and Gupta 2011). Scenario 9 is excluded from the discharge analysis because reasonable discharge cannot be simulated, as most discharge is generated through drainage flow and base flow at the Skjern catchment.

3. Spatial performance metrics

In this study, the baseline model is defined as a reference to benchmark spatial similarity of LST and ET patterns simulated by the scenarios. The spatial patterns simulated by the scenarios, denoted as scen, are

compared with the baseline model, denoted as ref, at each day between 2008 and 2010, which allows us to investigate seasonal variability in spatial sensitivity. To reduce the effect of small-scale variability and to investigate the predominant spatial patterns instead, the simulated spatial patterns are upscaled from their native 500-m resolution to 1 km.

a. Variogram

Variogram analysis is a frequently applied geostatistical tool that analyses the spatial autocorrelation structure of spatial data as a function of distance (lag; Deutsch and Journel 1998). Several studies have incorporated variograms to assess spatial patterns of soil moisture at catchment scale (Rosenbaum et al. 2012; Western et al. 1998). Furthermore, Koch et al. (2016a) and Korres et al. (2015) incorporate a variogram analysis of observed and simulated soil moisture patterns to assess the capability of integrated catchment models to represent the observed spatial patterns. Besides these previous efforts, the variogram analysis is typically not applied on spatial patterns of other land surface variables, such as ET or LST.

The omnidirectional empirical semivariance γ of variable Z at lag h is calculated as

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} (Z_i - Z_{i+h})^2, \qquad (7)$$

where $n(h)$ is the total number of data pairs at a given lag h, Z_i represents the value of Z at location i, and Z_{i+h} is the value separated by lag h from i. In the Skjern catchment, γ is calculated at eight irregularly spaced lags from 1 up to 25 km, which is about half of the maximum possible lag distance in the catchment. To quantify agreement between baseline $[\gamma(h)_{\mathrm{ref}}]$ and scenario $[\gamma(h)_{\mathrm{scen}}]$ semivariances, the root-mean-square error ($\mathrm{RMSE_{Vario}}$) is computed at the eight given lags:

$$\mathrm{RMSE_{Vario}} = \sqrt{\frac{\sum_{h=1}^{8} [\gamma(h)_{\mathrm{ref}} - \gamma(h)_{\mathrm{scen}}]^2}{8}}. \qquad (8)$$

In summary, the variogram metric focuses on the spatial autocorrelation as a function of distance and states a global measure of spatial performance that is not constrained by local agreement. Furthermore, the metric is bias insensitive because it only considers the relative deviation between values.

b. Empirical orthogonal functions

The EOF analysis is often applied to gain insight into the spatiotemporal variability of hydrological states and fluxes, such as soil moisture (Graf et al. 2014; Perry and Niemann 2007) or LST (Mascaro et al. 2015). The main feature of the EOF analysis is its capability to decompose a large spatiotemporal dataset of a single hydrological variable into two insightful components: 1) a set of orthogonal EOF maps that are invariant in time and 2) a set of loadings that quantify the importance of each EOF over time. The first EOF captures the highest portion of variance and the following EOFs subsequently add to the amount of explained variance. A more detailed mathematical description of the EOF methodology is given by Perry and Niemann (2008). While most studies have applied the EOF analysis on either observational or modeled data, more recent studies successfully applied it as a tool to spatially validate distributed hydrological models at catchment scale (Fang et al. 2015; Koch et al. 2015). Koch et al. (2015) brought forward the novel concept of performing a joint EOF analysis on an integral data matrix that contains both reference and modeled data. This joint analysis guarantees that the obtained EOF maps honor the spatiotemporal variability of both datasets and allows that the deviation between the loadings at a certain time step can be utilized as an indicator of spatial similarity. Following this approach, the associated loadings of highly similar maps at a certain time step will show a minimal deviation. Vice versa, a large loading deviation can be attested to pattern dissimilarity. The loading deviation must be weighted according to the variance contribution of the corresponding EOF to ensure a reliable pattern skill score. Thus, the EOF-based similarity score S_{EOF} between a reference and a scenario map at time x can be formulated as

$$S_{\mathrm{EOF}}^x = \sum_{i=1}^{n} w_i |(\mathrm{load}_i^{\mathrm{ref}_x} - \mathrm{load}_i^{\mathrm{scen}_x})|, \qquad (9)$$

where w_i is the variance contribution of the ith EOF, $\mathrm{load}^{\mathrm{scen}}$ is the scenario loading, and $\mathrm{load}^{\mathrm{ref}}$ is the reference loading at time x. Prior to the EOF analysis, the mean is removed from each map; thus, the methodology is based on the spatial anomalies, which makes it bias insensitive.

c. Fractions skill score

The FSS was initially developed by Roberts and Lean (2008) to obtain a scale-dependent quantitative measure of spatial performance for various competing precipitation forecasts. This methodology is based on fractions that are calculated at each cell and relate to occurrences of values exceeding a certain threshold at a given window size (scale). In most FSS applications, the

thresholds represent a few selected percentiles of interest, which has the purpose to eliminate any impact of a potential bias. As an example, Roberts (2008) analyzed precipitation forecasts at different extreme percentiles (75th, 90th, 95th, and 99th). To our knowledge, FSS has exclusively been applied on precipitation forecasts (e.g., Gilleland et al. 2009; Mittermaier et al. 2013; Wolff et al. 2014), but the methodology can easily be adapted to any continuous spatial data, such as ET or LST. In summary, FSS can be described via the following steps: 1) for each threshold of interest, truncate the reference (ref) and the scenario (scen) spatial patterns into binary maps; 2) for each cell, compute fractions within a given scale n based on the number of cells that exceed the threshold and lie within a window of size $n \times n$; and 3) calculate the mean-square error (MSE) between the reference and scenario fractions and normalize it with a worst-case MSE that reflects the condition with zero agreement between ref and scen (MSE_{wc}). The MSE is based on all cells N_{xy} that lie within the modeling domain with dimension of N_x and N_y. For a certain threshold, FSS at scale n is given by

$$\text{FSS}_{(n)} = 1 - \frac{\text{MSE}_{(n)}}{\text{MSE}_{(n)\text{wc}}}, \tag{10}$$

where

$$\text{MSE}_{(n)} = \frac{1}{N_{xy}} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [\text{ref}_{(n)ij} - \text{scen}_{(n)ij}]^2 \tag{11}$$

and

$$\text{MSE}_{(n)\text{wc}} = \frac{1}{N_{xy}} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \text{ref}_{(n)ij}^2 + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \text{scen}_{(n)ij}^2 \right]. \tag{12}$$

FSS ranges from zero to one, where one indicates a perfect agreement between reference and scenario patterns and zero reflects the worst possible performance. The methodology is bias insensitive because of the percentile truncation, which is a favorable feature for our application. Figure 2 illustrates the essential FSS steps for a summer ET pattern at the Skjern catchment. The 80th percentile is chosen to address the highest 20% of cells, and fractions at two different scales (5 and 10 km) underline the flexibility of this methodology. Adjusting the scale allows the user to tolerate placement errors, and thus uncertainties of the pattern, in dependence of the threshold. For instance, it may be desirable to accept higher uncertainties in the top 1% than in the top 25% of ET cells.

d. Connectivity

In hydrogeology, connectivity is a common measure to quantify aquifer heterogeneity (dell'Arciprete et al. 2012; Koch et al. 2014; Renard et al. 2011), where the degree of connectivity has direct physical implications on groundwater flow and solute transport. Only a few studies have implemented the concept of connectivity to describe spatial patterns of land surface variables. In this context, Western et al. (2001) and Grayson et al. (2002) analyzed connectivity of soil moisture patterns at the small Tarrawarra research catchment in Australia and linked it to discharge dynamics. These studies underline the physical relevance of connectivity as a characteristic of spatial patterns of hydrological states and fluxes at the land surface. Recently, Koch et al. (2016b) implemented a connectivity analysis in a spatial validation study of continental-scale land surface models using remotely sensed LST. Connectivity provided a reliable measure of heterogeneity and the general structure of the spatial LST patterns, which helped to identify specific shortcomings of the spatial performance. Renard and Allard (2013) suggested the following steps to conduct a connectivity analysis of a continuous variable: 1) decompose the domain into a series of binary maps using a stepwise increasing threshold that truncates the data along a range of selected thresholds that are typically defined by the variable's percentiles, 2) perform a cluster analysis on each of the binary maps to identify connected clusters, and 3) describe the transition from many disconnected clusters to a single connected cluster as the threshold increases. The latter can be quantified with principles of percolation theory; Hovadik and Larue (2007) proposed the probability of connection as a suitable metric to quantify how percolated clusters are. This metric, denoted as $\Gamma(t)$, is computed for each threshold t and describes the proportion of pairs of cells that are connected among all possible pairs of connected cells:

$$\Gamma(t) = \frac{1}{n_t^2} \sum_{i=1}^{N(X_t)} n_i^2, \tag{13}$$

where n_t is the total number of cells in the binary map X_t below or above threshold t, which has $N(X_t)$ distinct clusters in total, and n_i is the number of cells in the ith cluster in X_t. The percolation is described through an increasing threshold that moves along all percentiles of the variable's range. Focusing on the percentiles makes this methodology bias insensitive, and it allows investigating the clusters of high and low values. In the following, the connectivity analysis is separated and applied individually on cells that exceed a given
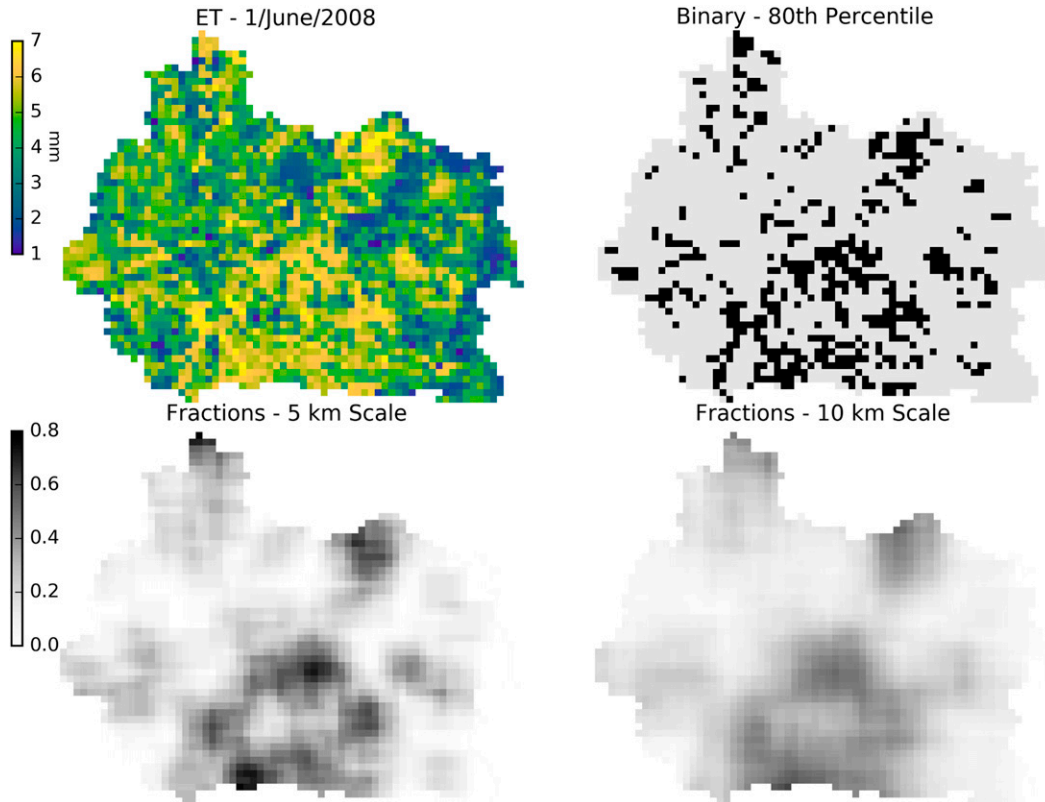
FIG. 2. (top left) The methodology of the FSS is illustrated in a map of simulated ET (mm day^{-1}) for a single day in June 2008. First, a percentile of interest is chosen and the continuous map is classified into a binary map by truncation at the given percentile. (top right) For this example, the 80th percentile is selected to display the top 20% of ET cells. The binary map is smoothed using (bottom left) 5 and (bottom right) 10 km, and the fractions are given for each cell.

threshold and those that fall below. The two approaches are referred to as high and low phase, respectively.

To compare the connectivity of a reference spatial pattern $\Gamma(t)_{\text{ref}}$ with a scenario $\Gamma(t)_{\text{scen}}$, the root-mean-square error between the reference and scenario connectivity $\Gamma(t)$ (RMSE$_{\text{CON}}$) can be computed for both phases:

$$\text{RMSE}_{\text{CON}} = \sqrt{\frac{\sum_{t=1}^{100} [\Gamma(t)_{\text{ref}} - \Gamma(t)_{\text{scen}}]^2}{100}}. \qquad (14)$$

The RMSE provides a global skill assessment of the connectivity that is not constrained by local agreement.

e. Multiple-point compatibility

In the field of geostatistics, multiple-point simulation algorithms are frequently applied frameworks to simulate heterogeneous geological structures with the aim to address geological uncertainty (He et al. 2013; Mariethoz et al. 2010). In comparison to conventional two-point simulations, by means of variograms or

transition probabilities (He et al. 2014) that are solely based on data, multiple-point simulations are informed by training images that incorporate the geologist's contextual insights as prior information (Mariethoz and Caers 2014; Mariethoz and Kelly 2011). Obtaining an accurate training image is crucial for the quality of a simulation, and Pérez et al. (2014) suggested a methodology to identify the most suitable training image by quantifying the spatial similarity of multiple competing training images with respect to reference data. The algorithm can easily be applied on spatial patterns of continuous land surface variables. For this study, the scenarios are considered as training images and the baseline model is set as reference. This metric is referred to as MPC in the coming sections. The basic idea behind MPC is to quantify spatial similarity as the proportion of spatial patterns that are found in both the reference and the simulated map. As indicated by Fig. 3, a spatial pattern is defined by a spiral search that starts at every cell that contains data and collects data from all cells it passes through. The number of data that are detected by the spiral search defines the pattern order. The user has
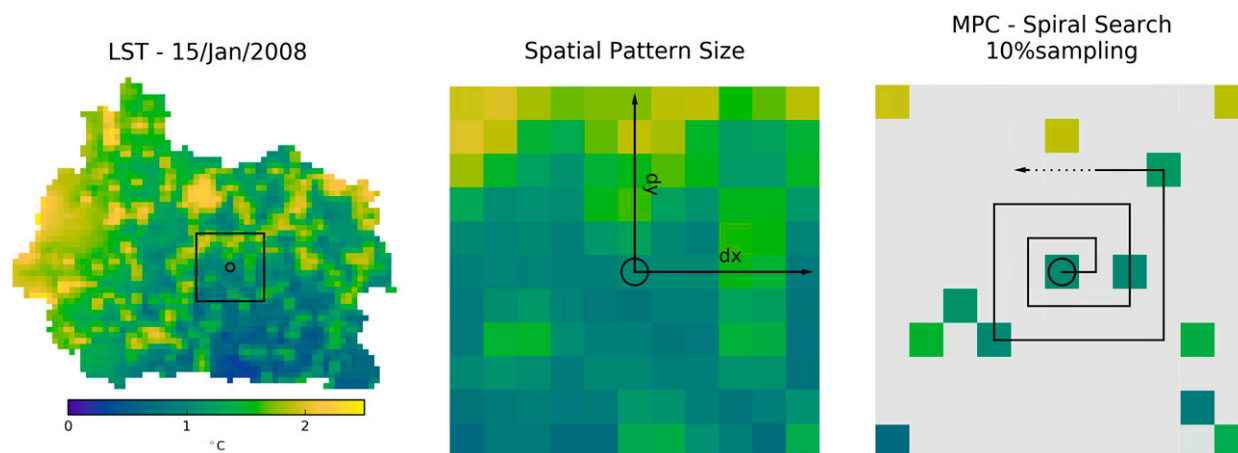
FIG. 3. (left) The spiral search that characterizes a spatial pattern in the MPC analysis is illustrated for a map of simulated LST in January 2008. (center) First, a window size (dy and dx) is selected that defines the size of the pattern, then (right) a spiral search registers all cells that contain data and the number of cells defines the order of the spatial pattern. The window is sampled with a certain percentage to allow investigating large patterns rather than focusing on small-scale variability.

to set a search radius (dxy) that defines the maximum extent of the spiral search and a sampling percentage that specifies how many data within dxy are recorded by the spiral search. The sampling option is included in the MPC algorithm to allow investigating large spatial patterns rather than focusing on small-scale spatial variability. After a spatial pattern in the reference is inventoried, the occurrence of this pattern is searched using a random search path across the scenario map. The pattern is checked off if the positions and values in the reference pattern are matched by a scenario pattern. Since it is difficult to find a perfect match, an error tolerance is introduced. To sidestep the bias sensitivity of MPC, we decided to transform the data to their spatial anomalies by removing the spatial mean at each time step. The metric ranges from zero to one and indicates the fraction of patterns contained in the reference that exist at any location in the simulation. Hence, MPC does not reflect any local agreement between simulation and reference. Instead, it focuses on the global assessment of the spatial structure. Last, the parameterization enables flexibility; for example, the search radius can easily be adjusted to investigate large and small patterns separately.

4. Results and discussion

a. Variogram

The omnidirectional semivariance is computed at eight irregularly spaced lag distances between 1 and 25 km. Figure 4 illustrates the resulting LST and ET variograms for the baseline model and the scenarios for a summer day in July 2009. The degree of spatial detail differs between the scenarios, which results in

different simulated spatial patterns that are equipped with individual variograms. Some scenarios have distinctly similar variograms in comparison to the baseline model, which suggests a limited spatial sensitivity. Others show a clear reduction in semivariance, indicating a decreased spatial variability and hence a high spatial sensitivity, such as the precipitation scenario for both variables and the groundwater coupling scenario for ET. However, some scenarios have a slightly increased semivariance with respect to the baseline, which seems counterintuitive at first, because the spatial detail is reduced in the scenarios. In general, homogenizing parameter or forcing fields poses complex transformations to the spatial patterns that are intractable and may result in a shift from energy limitation to water limitation or vice versa, which can lead to an increase of spatial variability.

b. EOF

The EOF analysis is computed based on daily spatial anomalies, and the integral spatiotemporal data matrix includes data from both scenario and baseline. The matrix dimensions are thus 2410×2192, where each row represents a position in space (2410 cells) and each column constitutes a daily time step from the 3-yr period (1096 days), which is doubled in order to accommodate data from both baseline and scenario. The resulting EOF maps and their associated loadings are presented in Fig. 5 for the assessment of simulated ET patterns of scenario 9 (disabled groundwater coupling). The first three EOFs are shown, which, as linear combination with their loadings, explain 64% of the variance. In general, multiplying the loadings with the EOF maps indicates how cells are related to the mean at a given
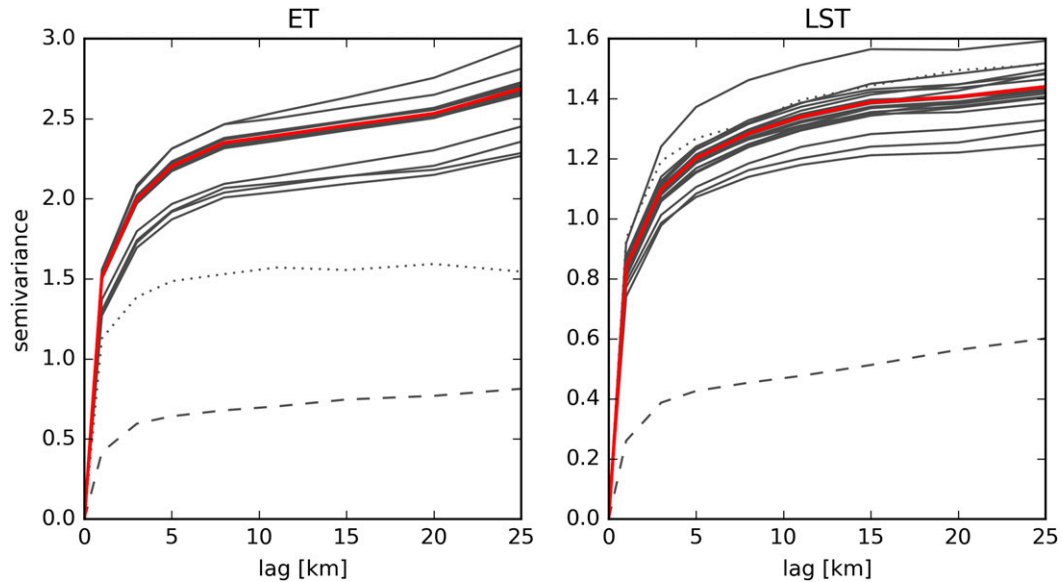
FIG. 4. Results from the variogram analysis for ET and LST patterns simulated by the baseline model (red) and all 22 scenarios (gray) on a single day (2 Jul 2009). Two scenarios are highlighted: scenario 6 (dashed) and scenario 9 (dotted). The omnidirectional semivariance is calculated at 10 lags spanning from 1 to 25 km.

time step. The predominant spatial pattern is captured by the first EOF, and it depicts the temporal evolution of the growing season with positive values for the agricultural cells that have a strong seasonal signal and negative values for the forest cells that show a dampened seasonal signal. The loadings indicate that scenario 9 comprises this pattern well, but honors it to a lesser degree because the loadings are generally lower in comparison to the baseline. Wintertime loadings are close to zero because little spatial variability is present in the ET patterns. The second EOF (EOF2) adds 21% to the explained variance and the loadings change sign, which indicates that this pattern is inverted during the season. The physical meaning behind EOF2 is less clear; however, the loadings of the baseline model and scenario 9 are almost identical, indicating that EOF2 captures the similarities between baseline and scenario ET patterns. Because of the spiky loadings time series, it can be speculated that EOF2 is related to the coupling of ET and precipitation. Furthermore, ET patterns from the baseline model and scenario 9 can be expected to be most similar during times when ET is controlled by precipitation, which diminishes the local groundwater influence. Last, the loadings of the third EOF (EOF3) are very divergent between the baseline model and scenario 9, and hence EOF3 captures major dissimilarities. Negative values in the EOF3 map reflect the critical zone where the water table is shallow and influences ET. The loadings of scenario 9 are positive and peak in the summertime, which underlines that ET is

clearly underestimated in those areas with respect to the baseline model. Following Eq. (9), the weighted loading deviation can be utilized as a metric to quantify spatial similarity.

c. FSS

The unique feature of FSS is its flexibility with respect to threshold and scale, and this is, to our knowledge, the first FSS application on hydrological land surface variables. Figure 6 depicts results for the evaluation of ET patterns for a summer day in June 2010 for three threshold percentiles: 1st, 5th, and 20th. At all thresholds FSS increases with scale, and FSS tends to one at large scales, which is an elemental characteristic because the number of cells in the binary maps for the baseline model and the scenarios are equal at each threshold. However, the rate of increase differs among the thresholds, because a natural higher degree of uncertainty is associated with the low percentiles in comparison to the higher percentiles. Overall, FSS clearly differentiates the scenarios with respect to their pattern similarity to the baseline model. To reduce computational time, the entire combination of scales and thresholds does not need to be calculated, and instead we suggest defining critical scales at which FSS is calculated. Uncertainty associated with the lowest 1% of cells may be tolerated better than higher percentiles. Therefore, the three percentiles are assessed at individual critical scales: 1st at 25 km, 5th at 15 km, and 20th at 5 km. These critical scales are subjective, but they
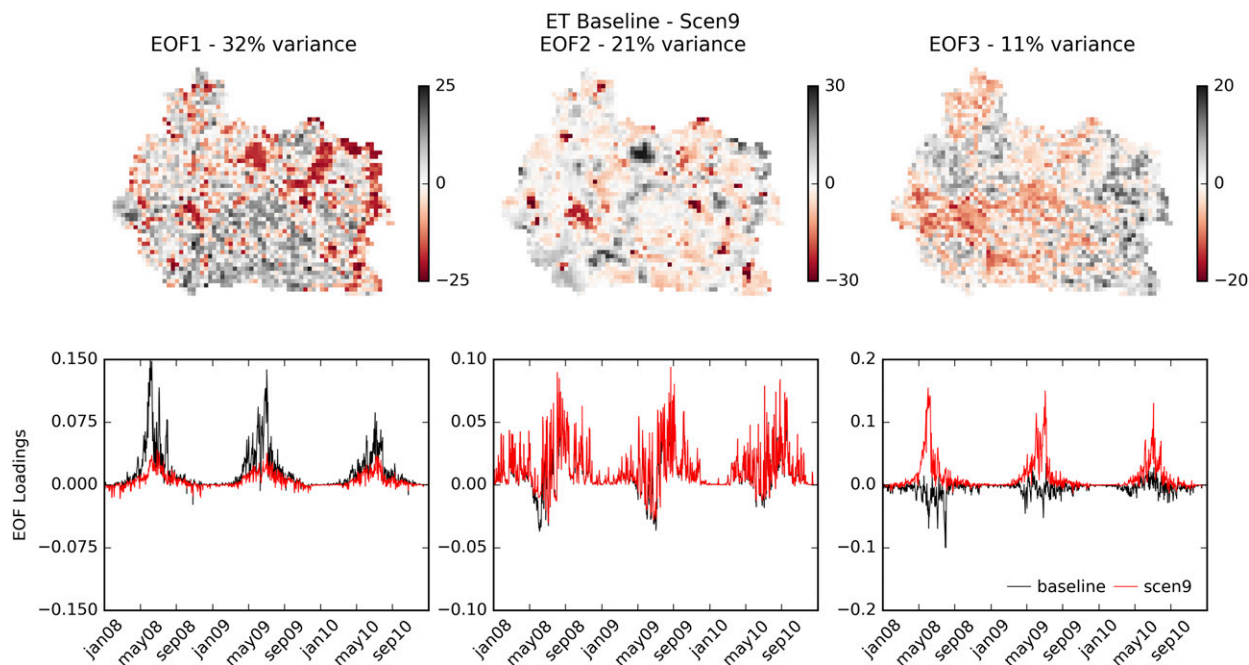
FIG. 5. Results from the EOF analysis comparing the spatial patterns of ET between the baseline model and scenario 9 (no groundwater coupling) for a 3-yr period (2008–10). (top) EOF maps and the amount of explained variance by each EOF. (bottom) The associated loadings for the baseline model and for scenario 9.

were found suitable for our application. This is conducted analogously for the three top percentiles—80th, 95th, and 99th—that focus on the highest 20%, 5%, and 1% of cells, respectively. The averages of the three top and bottom percentiles are calculated as an overall pattern similarity score and are referred to as "FSS high" and "FSS low" in the following analysis. One drawback of FSS is that focusing on percentiles may force spatial variability onto maps; for example, wintertime ET is small and spatial variability is marginal in the Skjern catchment, but classifying the map using percentiles artificially creates spatial variability. However, the percentile approach makes this methodology bias insensitive, which is favorable in our opinion.

d. Connectivity

The connectivity analysis provides insights into the spatial structure of a pattern and quantifies the degree of percolation as a function of threshold. The methodology is exemplified in Fig. 7 for LST patterns on a spring day in March 2009 for scenario 1 (no spatial variability in air temperature) and the baseline model. The cluster analysis at threshold percentiles is an elementary feature of the connectivity analysis, and Fig. 7 depicts the resulting cluster maps at two percentiles: the 20th and 80th percentiles that focus on the lowest 20% and highest 20% of cells, respectively. In general, LST is strongly coupled to

air temperature (Vancutsem et al. 2010), and the air temperature gradient in the Skjern catchment is characterized by warmer temperatures near the coastline in the west and a decrease in temperature toward east. Outside the growing season, where vegetation activity is minor, LST patterns exhibit a similar gradient. This results in a clear separation of the highest and lowest 20% of cells that form well-connected clusters in the baseline model, as indicated in Fig. 7. In scenario 1, this gradient is eliminated, which leads to a more scattered cluster allocation for the two given percentiles. Finally, the probability of connection at all thresholds can serve as a metric to quantify percolation as a function of threshold percentile. Abrupt increases in connectivity are distinct characteristics of connectivity curves and are unique attributes of spatial patterns (Renard and Allard 2013). The connectivity of both phases (low phase: cluster analysis for cells < threshold; high phase: cluster analysis for cells > threshold) is generally higher for the baseline model, which indicates that the spatial patterns are generally smoother and more homogeneous in comparison to scenario 1. As indicated by Eq. (14), the RMSE between the connectivity curves is incorporated as a metric to quantify the spatial performance for the low (CON low) and high phase (CON high) separately. Analogous to FSS, the connectivity
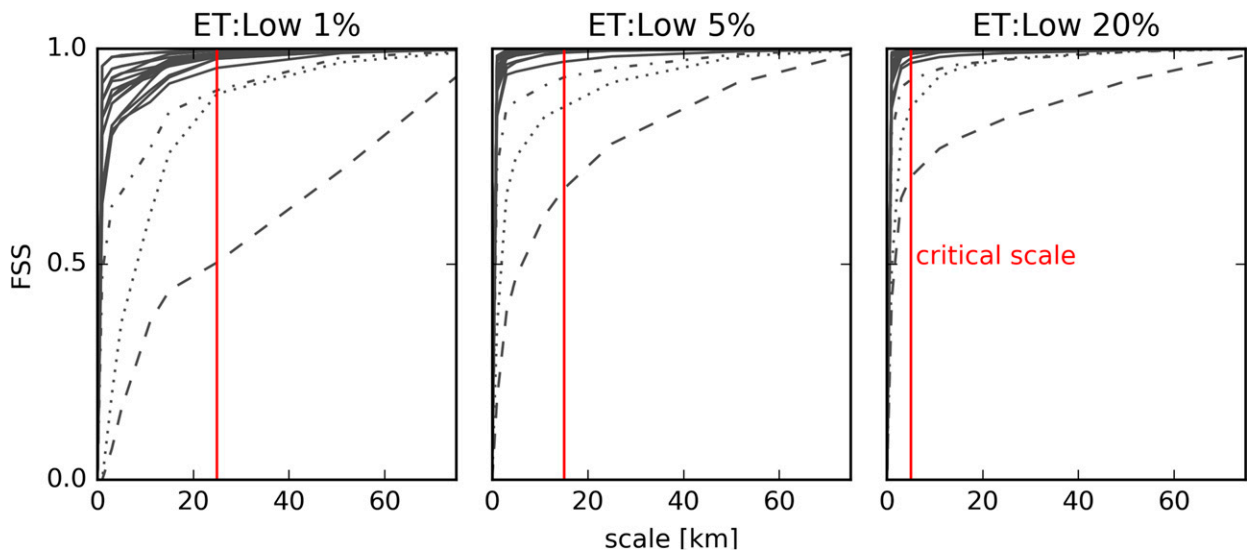
FIG. 6. The variability of FSS obtained from the comparison of ET patterns on 1 Jul 2010 between the baseline model and 22 scenarios. Each gray line represents the spatial performance of a single scenario, and the red lines mark the critical scales for each percentile that is used to calculate the merged FSS. The critical scales are 25 km for 1%, 15 km for 5%, and 5 km for 20%. Three scenarios are highlighted: scenario 6 (dashed), scenario 9 (dotted), and scenario 15 (dash–dotted).

analysis is based on threshold percentiles that may force spatial variability to patterns that are very homogeneous.

e. MPC

The MPC algorithm is very flexible because of its parameterization, but at the same time, care must be taken when defining the essential parameters, namely, pattern order, search radius, sampling percentage, and error threshold. Figure 8 reflects on this issue and plots the number of valid patterns as a function of the MPC parameters. An invalid pattern occurs if the spiral search does not reach the required pattern order within the given search radius, which is typically the case for cells close to the catchment boundary. These boundary effects are caused by the geometry of the catchment domain. The spiral search starts at each of the 2410 cells that comprise the Skjern catchment. With 100% sampling, ~1800 cells have valid patterns with full coverage when applying a 2-km search radius, which implies that the 5×5 search window contains 25 cells with data. The fully covered 5-km search radius that defines an 11×11 spiral search and has a pattern order of 121 is valid for ~1200 cells. The number of valid patterns increases for lower pattern orders, because empty cells are tolerated in the search window. Sampling reduces the maximum pattern order; for instance, 33% sampling of the 5-km search radius lowers the maximum pattern order from originally 121 (11×11 search window) to 40. This allows MPC to assess large patterns without focusing on small-

scale variability. For this study, we define two MPC parameterizations (indicated in gray in Fig. 8). The first investigates large patterns at a 5-km search radius with 10% sampling, and it considers pattern orders between 8 and 12. The second setup focuses on small patterns with a search radius of 2 km. Here the sampling percentage can be higher (20%), because small-scale variability is more desirable and pattern orders of 4 and 5 are taken into consideration. In the following the two described MPC parameterization schemes are referred to as "MPC large" and "MPC small." Furthermore, the error threshold, which is used for the comparison of inventoried patterns, has to be defined. The error threshold should be determined in dependence of the variable of interest. A straightforward and objective option is to draw the value from the variable's standard deviation. We define MPC large to tolerate an error of one standard deviation when comparing patterns of LST and ET, and MPC small is set up with an error threshold of half a standard deviation. This differentiation enables MPC large to address larger gradients that have a natural higher variability, whereas small patterns have lower gradients, and thus, accuracy has to be higher. The standard deviations for LST and ET are derived from the baseline model and are 0.62°C and 0.27 mm, respectively.

f. Seasonal sensitivity

The set of eight spatial performance metrics described above (variogram, EOF, FSS high, FSS low, CON low,
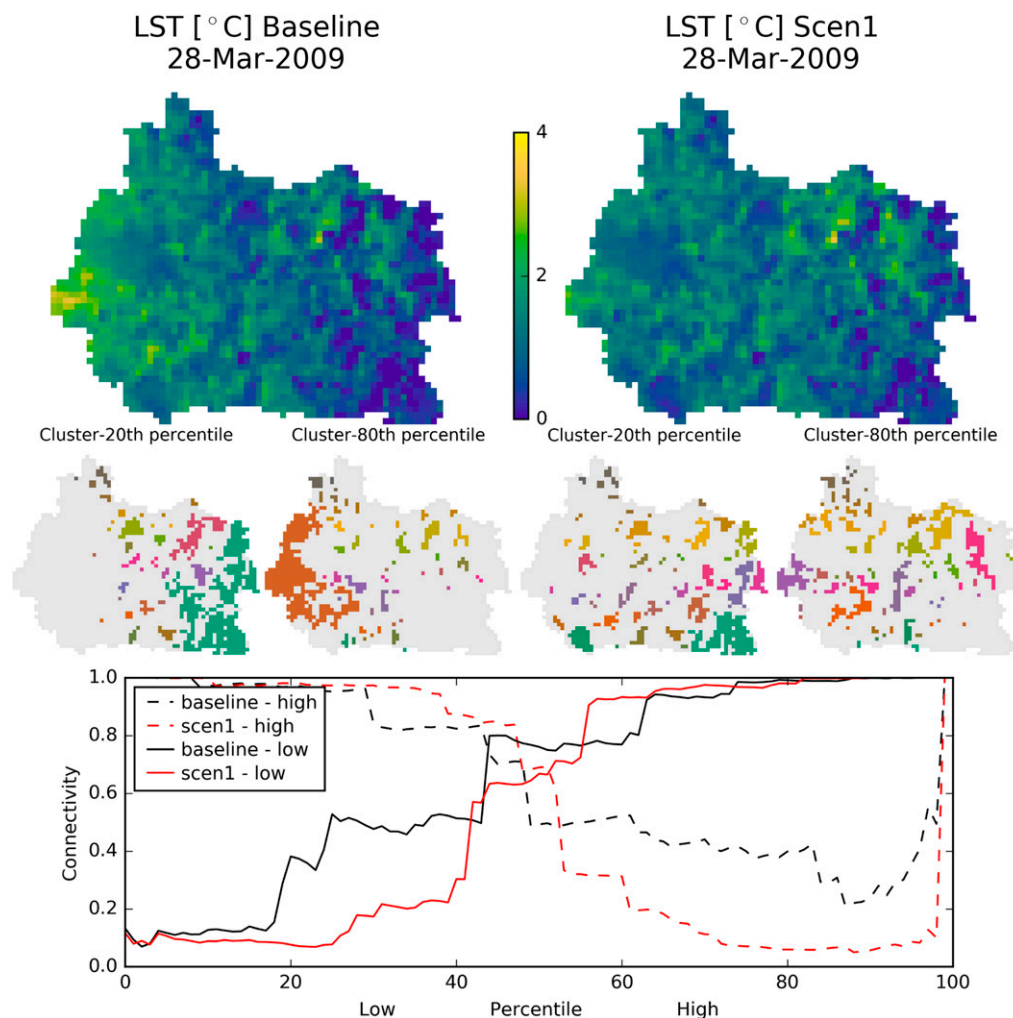
FIG. 7. (top) Connectivity analysis illustrated for a comparison of LST maps simulated on 28 Mar 2008 by (left) the baseline model and (right) scenario 1 (no spatial variability in air temperature). The continuous maps are first truncated at specific thresholds, and then a cluster analysis is performed on the basis of the binary maps. (middle) The results from the cluster analysis at the 20th and 80th percentiles. The first analyzes the coldest 20% of cells and the latter focuses on the warmest 20% of cells. A unique color is assigned to each connected cluster. (bottom) The computed connectivity metric is shown for the high and low phase at each percentile.

CON high, MPC small, and MPC large) is applied on 3 years of daily LST and ET patterns of the 22 scenarios, and spatial similarity is quantified with respect to the baseline model. As an example, Fig. 9 depicts the daily spatial performance based on the connectivity analysis of the high phase of LST patterns for two scenarios. Both scenarios 10 and 14 perturb the LAI input, where the first is based on a 10-km smoothing and the second homogenizes the LAI fields to spatially uniform values. The connectivity analysis identifies small changes in the spatial patterns during wintertime, but two peaks, one at the start and one at the end of the growing season, are detectable in all 3 years. These peaks underline the distinct seasonality in LST sensitivity to LAI. Furthermore, the

connectivity of the warm LST clusters shows similar scores for the late growing season in both scenarios. In this period, a homogeneous LAI input informs the model equally well as a 10-km smoothed LAI input. However, during the early growing season, the latter performs clearly better than scenario 14 with constant LAI. In summary, the spatial patterns of LST are most complex during the late cropping season, where information at the 10-km scale is clearly insufficient. During the middle of the growing season, where vegetation is active throughout the catchment, the connectivity of the scenarios is in good agreement with the baseline model, which supports the interpretation that spatial patterns are simple and detailed information on LAI is less crucial.
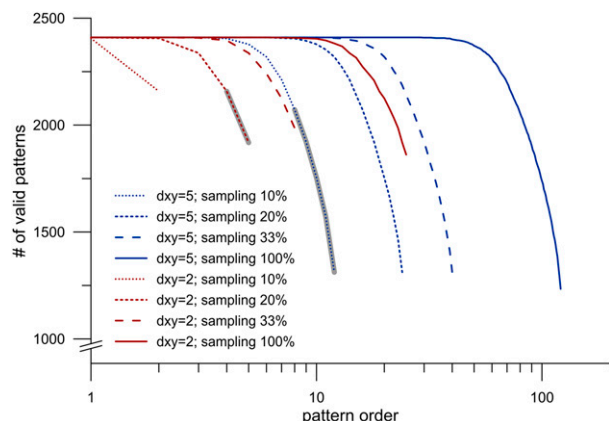
FIG. 8. The number of cells that contain data within a window of size dxy defines the order of the spatial pattern for the MPC analysis. Sampling is included in order to investigate large patterns rather than focusing on small-scale variability. The number of valid events represents the cells that ensure a certain pattern order given a window size (dxy) and a sampling percentage. The gray lines mark the range of order and the sampling that is chosen for this study to reflect small patterns ($dxy = 2$; sampling = 20%; orders 4 and 5) and large patterns ($dxy = 5$; sampling = 10%; orders 8–12).
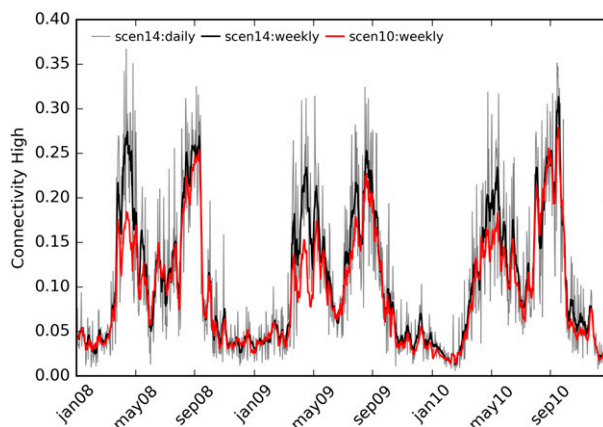

FIG. 9. The results obtained from the connectivity analysis of the high phase of LST is given for scenario 10 (10-km smoothed spatial variability of LAI) and scenario 14 (no spatial variability of LAI). Low values indicate spatial similarity. The metric is computed on the daily maps and, for better visualization, a 7-day smoothing filter is applied on the daily data.

To summarize the most important findings, Figs. 10 and 11 show the spatial performance of all eight metrics for the four most sensitive scenarios for ET and LST in a single year. The most sensitive scenarios are identified by simply selecting the ones that exhibit the lowest pattern performance on average for the 3-yr period. The results for the ET patterns in 2008 (Fig. 10) stress that the sensitivity of most scenarios underlies a distinct seasonality. Furthermore, the seasonality of sensitivities is partly rated similar by the spatial performance metrics. Scenarios 6 and 9 are related to groundwater coupling and precipitation, respectively, and their sensitivity is highest in early summer. This conclusion is clearly indicated by EOF and variogram, whereas the other metrics provide a more nuanced interpretation. This underlines the complexity of comparing spatial patterns, because different metrics are expected to provide different interpretations of the spatial similarity. Perturbations in LAI and vegetation height are considered in scenarios 14 and 17, respectively, and their sensitivity is highest toward the end of the growing season. Scenario 17 has a distinct pattern sensitivity in wintertime, as indicated by FSS and connectivity, because the year-round constant vegetation height of the forest is averaged out with all nonforest areas. This underlines that a detailed land-use map and hence the correct allocation of forest cells is essential to represent wintertime processes correctly. EOF, variogram, and MPC that are not based on percentile thresholds attest little sensitivity to the scenarios in wintertime, because

spatial variability of ET is low during the energy-limited months.

Figure 11 depicts results for LST sensitivity, and in comparison to ET, scenario 9 is the sole scenario that shows a clear seasonality that follows the growing season. Additionally, the spatial sensitivity of air temperature is given, and it is rated most significant during winter by FSS and connectivity, which stresses the importance of the east–west air temperature gradient outside the growing season. The other metrics, which are not dependent on percentiles, evaluate air temperature as sensitive as well, but attest less seasonality. Again, scenario 17 is most sensitive during wintertime, which underlines the importance of vegetation height as a driver of spatial variability of LST. The sensitivity of LAI is addressed by scenario 14 and, similar to ET, two peaks are identified by most of the metrics, with the first at the start and the second at the end of the growing season.

g. Spatial information content

Comparing spatial patterns is not a trivial task because many dimensions of spatial information have to be considered simultaneously. Typically, a spatial metric focuses on a limited number of dimensions, such as the overall global structure or the local agreement, as indicated by Table 4. Metrics will naturally give ambiguous results that should not challenge their reliability and legitimacy. Instead, diverging interpretations between metrics should be embraced, because this stresses that different dimensions of spatial information are covered. On the other hand, metrics with similar
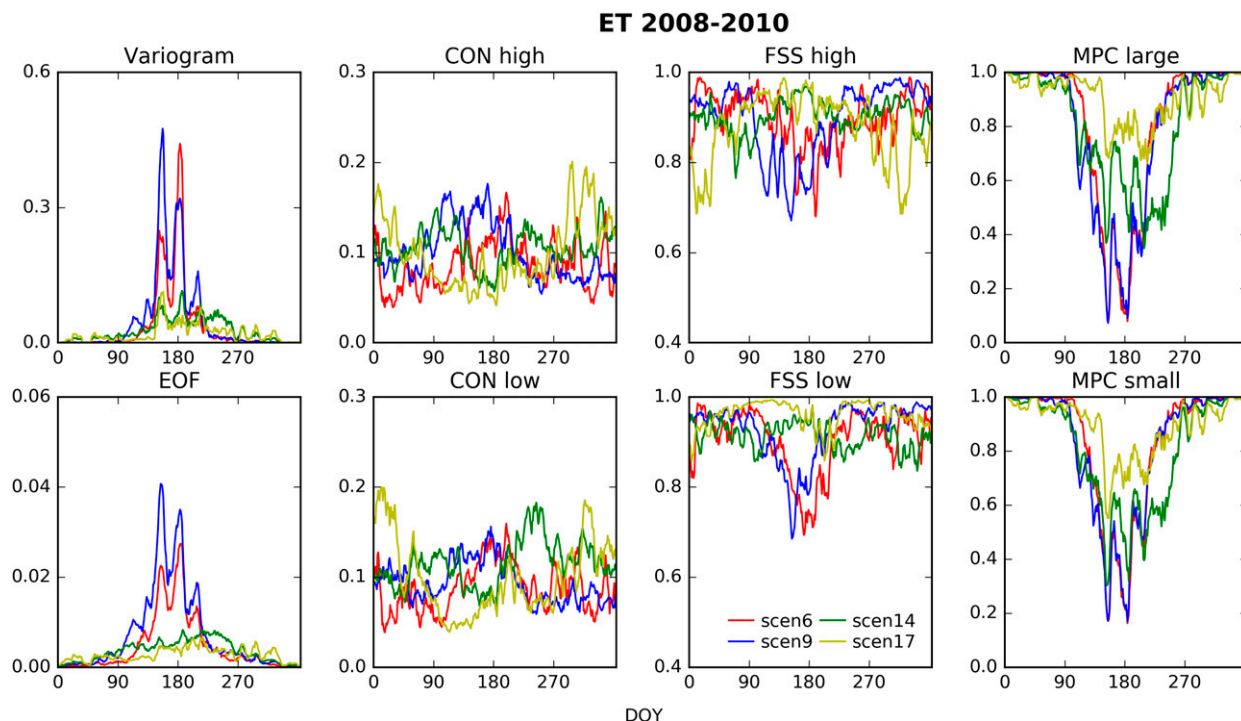
FIG. 10. Eight spatial performance metrics that assess the spatial similarity between the baseline model and four scenarios are shown for simulated ET patterns. The temporal variability of spatial similarity is given as averages of the smoothed metric scores for a 3-yr period (2008–10): scenario 6 (no spatial variability of precipitation), scenario 9 (no groundwater coupling), scenario 14 (no spatial variability of LAI), and scenario 17 (no spatial variability of vegetation height).

scores can be considered redundant. To analyze the uniqueness and redundancy of the set of applied performance metrics, Table 5 states the coefficient of determination R^2 between the metrics based on individual daily pattern similarity scores for ET and LST of all scenarios. The spatial correlation coefficient R, the overall bias, and the spatial RMSE are added in order to include some traditional and easy-to-compute metrics, which are based on cell-to-cell comparisons. Overall, the differences between the R^2 scores for the LST and ET patterns are marginal and hence interpretations are analogous. The bias explains negligible amounts of variance of the other metrics, which is logical given the bias insensitivity of the other metrics. To better understand drivers of spatial variability, the bias can be a relevant metric; for example, scenario 9, which decouples groundwater and land surface, results in a negative bias of up to $2 \, \text{mm} \, \text{day}^{-1}$ (not shown) due to the reduced water availability. However, we focus on bias-insensitive metrics and suggest applying these in combination with the bias when validating or calibrating a model spatially. The separation is crucial, because a metric that is sensitive to both the spatial pattern and the bias may be ambivalent and misleading. From a modeling perspective, the spatial pattern and

the bias are typically controlled by different parameters, and model diagnostics are most straightforward if metrics treat them separately.

The pattern scores from the spatial correlation coefficient are in reasonable agreement with FSS and the connectivity analysis because all are controlled by the tails of the distribution. Variogram, EOF, MPC, and RMSE are somewhat dependent on the spatial variability of the patterns, which results in high R^2 scores. Overall, part of the information of more complex spatial performance metrics can be covered by simpler cell-to-cell metrics. Based on the results of Table 5, the EOF analysis has the highest average R^2, which allows the conclusion that it is the most suitable standalone metric for the LST and ET patterns of the Skjern catchment. Connectivity analysis and FSS show least redundancy to the remaining metrics, which underlines that assessing spatial patterns at threshold percentiles clearly adds information to a pattern comparison. MPC is related to the variogram analysis because both assess a pattern via a sequence of cells: MPC takes multiple locations into consideration, whereas variogram analysis focuses on two locations. The differentiation of MPC into MPC small and MPC large provided limited additional information content to the
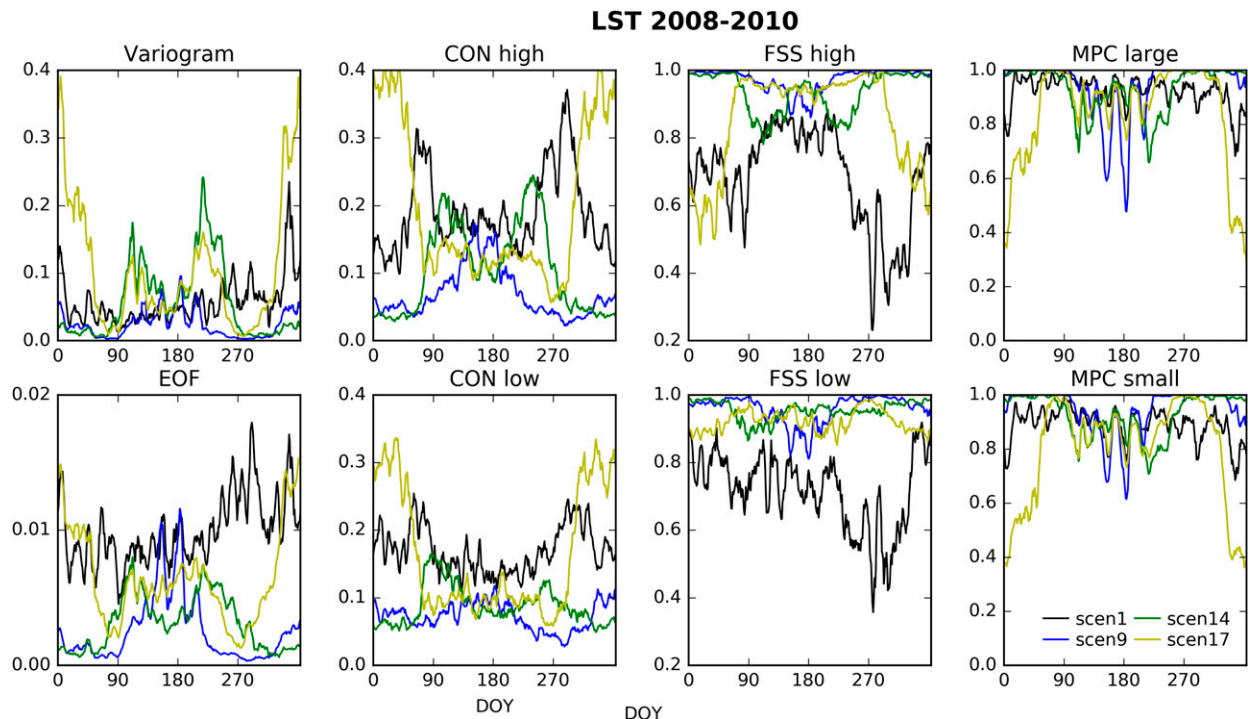
LST 2008-2010

FIG. 11. Eight spatial performance metrics that assess the spatial similarity between the baseline model and four scenarios are shown for simulated LST patterns. The temporal variability of spatial similarity is given as averages of the smoothed metric scores for a 3-yr period (2008–10): scenario 1 (no spatial variability of air temperature), scenario 9 (no groundwater coupling), scenario 14 (no spatial variability of LAI), and scenario 17 (no spatial variability of vegetation height).

pattern assessment. The connectivity of the low phase and the high phase are attributed with R^2 values of 0.55 and 0.66 for ET and LST, respectively, and hence they are not fully independent. However, the differentiation of FSS into FSS high and FSS low proves to be very reasonable, because they exhibit low R^2 scores.

h. Relative sensitivity

To summarize the results of all 11 spatial performance metrics, Fig. 12 depicts the relative sensitivity of each

scenario with respect to spatial patterns of ET and LST. The relative sensitivity is computed based on the average spatial similarity score over the entire 3 years. For each metric, the scenario with the lowest score is marked as most sensitive and the scenario with the best score is rated as the least sensitive scenario. In the following the relative sensitivity is linearly scaled between zero and one, where one defines the highest spatial sensitivity. The general agreement between the spatial performance metrics is striking. The only peculiar metric is the bias, which is not surprising, given that the other metrics

TABLE 4. Overview of the spatial performance metrics.

Metric	Focus	Data transformation	Data	Bias sensitive	Scale flexibility	Spatial validity	Reference
EOF	Predominant patterns and their temporal weighting	Spatial anomalies	Continuous	No	No	Local	Koch et al. (2015)
CON	Percolation of cluster maps at various thresholds	Categorize at thresholds	Binary	No	No	Global	Koch et al. (2016b)
FSS	Spatial agreement at various scales and thresholds	Categorize at thresholds	Binary	No	Yes	Local	Roberts and Lean (2008)
MPC	Higher-order spatial structure	Not necessary[a]	Continuous	Yes[a]	Yes	Global	Pérez et al. (2014)
Variogram	Spatial structure based on two-point variability	Not necessary	Continuous	No	Yes	Global	Korres et al. (2015)

[a] Data are transformed to spatial anomalies in the preprocessing of MPC to ensure bias insensitivity.

TABLE 5. The coefficient of determination between the applied spatial performance metrics. The analysis is based on the daily spatial similarity scores for a 3-yr period (2008–10) for all scenarios. Values above 0.5 are given in boldface to visualize significant dependencies. High R^2 indicates redundant information content between the assessed metrics.

R^2	Spatial R	Bias	RMSE	Variogram	EOF	CON high	CON low	FSS high	FSS low	MPC large	MPC small
ET											
Spatial R	1										
Bias	0.04	1									
RMSE	0.36	0.03	1								
Variogram	0.18	0.06	**0.53**	1							
EOF	0.48	0.09	**0.84**	**0.65**	1						
CON high	0.44	0.00	0.18	0.05	0.19	1					
CON low	0.44	0.00	0.21	0.07	0.22	**0.55**	1				
FSS high	**0.53**	0.02	0.17	0.07	0.21	0.43	0.32	1			
FSS low	**0.53**	0.00	0.31	0.12	0.34	0.26	0.39	0.19	1		
MPC large	0.24	0.01	**0.71**	**0.50**	**0.68**	0.13	0.17	0.12	0.22	1	
MPC small	0.18	0.01	**0.65**	**0.51**	**0.62**	0.11	0.14	0.10	0.16	**0.92**	1
LST											
Spatial R	1										
Bias	0.00	1									
RMSE	0.40	0.00	1								
Variogram	0.20	0.00	**0.55**	1							
EOF	0.48	0.00	**0.74**	**0.58**	1						
CON high	0.47	0.01	0.47	0.37	**0.59**	1					
CON low	0.42	0.01	0.39	0.28	**0.53**	**0.66**	1				
FSS high	**0.74**	0.00	0.30	0.18	0.47	0.47	0.43	1			
FSS low	**0.51**	0.00	0.23	0.05	0.34	0.26	0.31	0.41	1		
MPC large	0.22	0.00	**0.59**	**0.69**	**0.51**	0.25	0.18	0.15	0.04	1	
MPC small	0.27	0.01	**0.63**	**0.74**	**0.62**	0.39	0.30	0.23	0.07	**0.86**	1

are bias insensitive. Differences between spatial drivers of ET and LST are clear. Vegetation, groundwater, and precipitation are the most dominant drivers of ET patterns, whereas LST is most sensitive to climate forcing and vegetation height (scenario 17). It is unexpected to see low spatial sensitivity attested to albedo (scenario 16) and root depth (scenario 15) for the simulation of the two given land surface variables, because these are frequently used for calibrating integrated catchment model against discharge. Consequently, albedo and root depth are discharge sensitive because they are affecting the water balance, but their pattern sensitivity appears limited. The vegetation parameters derived from the land-use map (scenarios 19–22) are more sensitive to ET patterns than LST patterns. In particular, interception is a sensitive parameter to the simulated ET patterns. This is logical from a modeling perspective, because it is the first storage term to be evaporated, before transpiration or soil evaporation is initiated. Interestingly, scenarios 7 and 8, which are related to the soil type map and the hydrogeological model of the near surface, respectively, are mostly rated insensitive to the spatial patterns of LST and ET. These model parameters are typically included in calibration frameworks of catchment models against discharge, but our spatial sensitivity analysis stresses that soil physical parameters and hydraulic conductivity

have a limited sensitivity to energy balance–related variables at the land surface. The soils in the Skjern catchment are predominantly of sandy origin and variability in soil texture is less pronounced, which may cause the reduced relative sensitivity of scenarios 7 and 8.

Whether state-of-the-art or not, a model will always depict a simplified version of the natural systems. Therefore, the identified spatial sensitivities are not directly transferrable to the natural system. Furthermore, it is questionable if the conclusions are applicable to other catchments that are equipped with different characteristics.

5. Conclusions

This study provides a comprehensive scenario-based spatial sensitivity analysis of a highly parameterized catchment model (MIKE SHE) that is coupled to a land surface model (SW-ET). The application and rigorous testing of the set of innovative spatial performance metrics constitutes the key novelty of this study. The number of sophisticated, spatially distributed models is growing, and supported by the grand availability of spatial observations, the community requires methods to assess simulated spatial patterns. Further, the study aims to increase process understanding and to
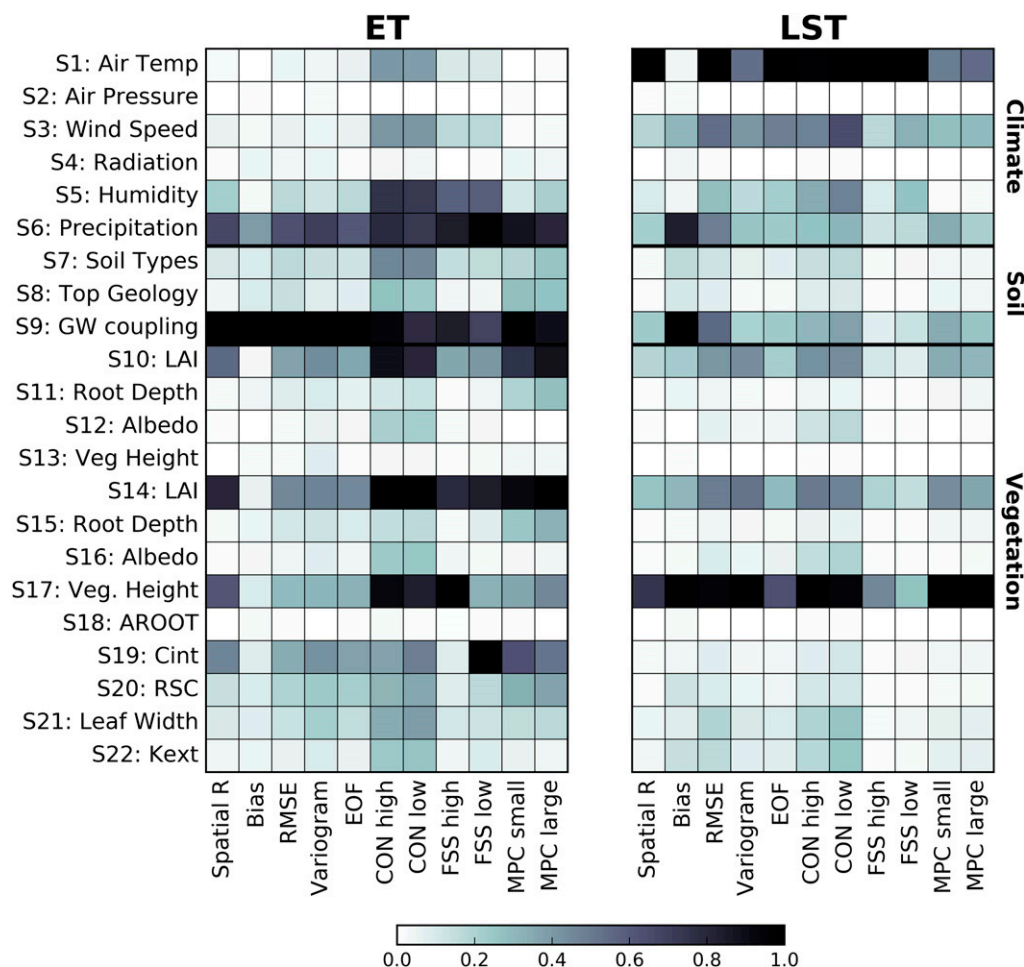
FIG. 12. The relative sensitivity is given for each metric for the respective scenarios. The sensitivity is linearly normalized relative to the most sensitive scenario (1.0) and the least sensitive scenario (0.0). The analysis is based on mean values for each scenario based on 3 years (2008–09).

enhance knowledge on the drivers of spatial variability of simulated land surface variables such as evapotranspiration (ET) and land surface temperature (LST). We draw the following main conclusions from this work.

- *Spatial performance metrics.* There is no unique answer to the quantification of similarity between spatial patterns, because such a comparison is based on many dimensions of spatial information. Each metric covers a limited number of these dimensions, and in the end, a combination of metrics may be the most applicable option for a reliable pattern comparison. This study features five innovative spatial performance metrics that are bias insensitive and should be considered in the standard toolbox of spatial verification methods in environmental modeling. True pattern metrics are inevitably needed to improve

models by diagnosing spatial deficiencies. Based on the analysis of ET and LST patterns of the Skjern catchment, the EOF analysis is the best option for a stand-alone metric. FSS clearly provides the most unique information content, and we introduced the concept of critical scales that can be specified by the user to tolerate threshold-dependent placement. Thus, a combination of EOF and FSS would be most appropriate for the analysis of ET and LST patterns at the Skjern catchment.

- *Scenario sensitivity.* Drivers of simulated spatial patterns of ET and LST could be concluded from the scenario-based spatial sensitivity analysis, and we found that the sensitivity of most scenarios underlies a distinct seasonality that often follows the growing season. Overall, vegetation, groundwater, and precipitation are the most important drivers of ET, while LST is predominantly driven by climate

forcing and vegetation height. In terms of ET patterns, spatial detail in the vegetation parameterization is most essential in the late growing season, where a 10-km smoothed LAI input provides equally poor results as having no spatial information for LAI. The groundwater coupling is more sensitive to ET than LST, and it is highest during summertime for the critical zone. LST is strongly correlated to air temperature, and the east–west gradient that is evident in the Skjern catchment is most prominent outside the growing season, where the effect of vegetation on LST is lowest. The correct allocation of forest cells is of high importance when defining the vegetation height, and sensitivity is highest for LST patterns during the winter. The spatial sensitivity of precipitation is stronger for ET than LST and clearly peaks in the early growing season.

This study presents a framework that focuses on spatial patterns that are often neglected in hydrological simulations. The findings regarding the spatial sensitivities may not be a universal contribution to the modeling community, but the proposed framework is certainly generic.

Hydrological states and fluxes at the land surface are driven by a complex interaction and feedbacks of climatic drivers and surface and subsurface parameterization. Further efforts should be undertaken to increase the spatial detail of model input, for example, by means of remote sensing to improve spatial predictability. Spatial observations of LST are widely available through various remote sensing platforms and should be included in the calibration and validation of complex distributed models. For such tasks, spatial performance metrics are essential to conduct a reliable assessment of the simulated spatial patterns.

REFERENCES

Abbott, M. B., J. C. Bathurst, J. A. Cunge, P. E. O'Connell, and J. Rasmussen, 1986: An introduction to the European Hydrological System—Système Hydrologique Européen, "SHE", 2: Structure of a physically-based, distributed modelling system. *J. Hydrol.*, **87**, 61–77, doi:10.1016/0022-1694(86)90115-0.

Andersen, J., G. Dybkjaer, K. H. Jensen, J. C. Refsgaard, and K. Rasmussen, 2002: Use of remotely sensed precipitation and leaf area index in a distributed hydrological model. *J. Hydrol.*, **264**, 34–50, doi:10.1016/S0022-1694(02)00046-X.

Bayer, P., and M. Finkel, 2007: Optimization of concentration control by evolution strategies: Formulation, application, and assessment of remedial solutions. *Water Resour. Res.*, **43**, W02410, doi:10.1029/2005WR004753.

Bierkens, M. F. P., and Coauthors, 2015: Hyper-resolution global hydrological modelling: What is next? *Hydrol. Processes*, **29**, 310–320, doi:10.1002/hyp.10391.

Bircher, S., N. Skou, K. H. Jensen, J. P. Walker, and L. Rasmussen, 2012: A soil moisture and temperature network for SMOS validation in western Denmark. *Hydrol. Earth Syst. Sci.*, **16**, 1445–1463, doi:10.5194/hess-16-1445-2012.

Boegh, E., and Coauthors, 2004: Incorporating remote sensing data in physically based distributed agro-hydrological modelling. *J. Hydrol.*, **287**, 279–299, doi:10.1016/j.jhydrol.2003.10.018.

Butts, M., and Coauthors, 2014: Embedding complex hydrology in the regional climate system—Dynamic coupling across different modelling domains. *Adv. Water Resour.*, **74**, 166–184, doi:10.1016/j.advwatres.2014.09.004.

Chaney, N. W., J. K. Roundy, J. E. Herrera-Estrada, and E. F. Wood, 2015: High-resolution modeling of the spatial heterogeneity of soil moisture: Applications in network design. *Water Resour. Res.*, **51**, 619–638, doi:10.1002/2013WR014964.

Corbari, C., and M. Marco, 2014: Calibration and validation of a distributed energy–water balance model using satellite data of land surface temperature and ground discharge measurements. *J. Hydrometeor.*, **15**, 376–392, doi:10.1175/JHM-D-12-0173.1.

Chiles, J.-P., and P. Delfiner, 2009: *Geostatistics: Modeling Spatial Uncertainty.* Wiley, 726 pp.

Clark, M. P., and Coauthors, 2015: A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resour. Res.*, **51**, 2498–2514, doi:10.1002/2015WR017198.

dell'Arciprete, D., R. Bersezio, F. Felletti, M. Giudici, A. Comunian, and P. Renard, 2012: Comparison of three geostatistical methods for hydrofacies simulation: A test on alluvial sediments. *Hydrogeol. J.*, **20**, 299–311, doi:10.1007/s10040-011-0808-0.

Deutsch, C. V., and A. G. Journel, 1998: *GSLIB: Geostatistical Software Library and User's Guide.* Oxford University Press, 340 pp.

Doherty, J., 2016: PEST: Model-independent Parameter Estimation: User Manual Part I: PEST, SENSAN, and global optimisers. 6th ed. Watermark Numerical Computing, 390 pp. [Available online at http://www.pesthomepage.org/getfiles.php?file=newpestman1.pdf.]

Fang, Z., H. Bogena, S. Kollet, J. Koch, and H. Vereecken, 2015: Spatio-temporal validation of long-term 3D hydrological simulations of a forested catchment using empirical orthogonal functions and wavelet coherence analysis. *J. Hydrol.*, **529**, 1754–1767, doi:10.1016/j.jhydrol.2015.08.011.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, doi:10.1175/2009WAF2222269.1.

Graf, A., H. R. Bogena, C. Drüe, H. Hardelauf, T. Pütz, G. Heinemann, and H. Vereecken, 2014: Spatiotemporal relations between water budget components and soil water content in a forested tributary catchment. *Water Resour. Res.*, **50**, 4837–4857, doi:10.1002/2013WR014516.

Grayson, R. B., G. Blöschl, A. W. Western, and T. A. McMahon, 2002: Advances in the use of observed spatial patterns of catchment hydrological response. *Adv. Water Resour.*, **25**, 1313–1334, doi:10.1016/S0309-1708(02)00060-X.

Hansen, N., and A. Ostermeier, 2001: Completely derandomized self-adaptation in evolution strategies. *Evol. Comput.*, **9**, 159–195, doi:10.1162/106365601750190398.

He, X., T. O. Sonnenborg, F. Jørgensen, A. S. Høyer, R. R. Møller, and K. H. Jensen, 2013: Analyzing the effects of geological and parameter uncertainty on prediction of groundwater head and travel time. *Hydrol. Earth Syst. Sci.*, **17**, 3245–3260, doi:10.5194/hess-17-3245-2013.

——, J. Koch, T. O. Sonnenborg, F. Jørgensen, C. Schamper, and J. C. Refsgaard, 2014: Transition probability-based stochastic geological modeling using airborne geophysical data and borehole data. *Water Resour. Res.*, **50**, 3147–3169, doi:10.1002/2013WR014593.

Hovadik, J. M., and D. K. Larue, 2007: Static characterizations of reservoirs: Refining the concepts of connectivity and continuity. *Petrol. Geosci.*, **13**, 195–211, doi:10.1144/1354-079305-697.

Jensen, K. H., and T. H. Illangasekare, 2011: HOBE: A hydrological observatory. *Vadose Zone J.*, **10**, 1–7, doi:10.2136/vzj2011.0006.

Jönsson, P., and L. Eklundh, 2002: Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE Trans. Geosci. Remote Sens.*, **40**, 1824–1832, doi:10.1109/TGRS.2002.802519.

——, and ——, 2004: TIMESAT—A program for analyzing time-series of satellite sensor data. *Comput. Geosci.*, **30**, 833–845, doi:10.1016/j.cageo.2004.05.006.

Koch, J., X. He, K. H. Jensen, and J. C. Refsgaard, 2014: Challenges in conditioning a stochastic geological model of a heterogeneous glacial aquifer to a comprehensive soft data set. *Hydrol. Earth Syst. Sci.*, **18**, 2907–2923, doi:10.5194/hess-18-2907-2014.

——, K. H. Jensen, and S. Stisen, 2015: Toward a true spatial model evaluation in distributed hydrological modeling: Kappa statistics, Fuzzy theory, and EOF-analysis benchmarked by the human perception and evaluated against a modeling case study. *Water Resour. Res.*, **51**, 1225–1246, doi:10.1002/2014WR016607.

——, T. Cornelissen, Z. Fang, H. Bogena, B. Diekkrüger, S. Kollet, and S. Stisen, 2016a: Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment. *J. Hydrol.*, **533**, 234–249, doi:10.1016/j.jhydrol.2015.12.002.

——, A. Siemann, S. Stisen, and J. Sheffield, 2016b: Spatial validation of large-scale land surface models against monthly land surface temperature patterns using innovative performance metrics. *J. Geophys. Res. Atmos.*, **121**, 5430–5452, doi:10.1002/2015JD024482.

Korres, W., and Coauthors, 2015: Spatio-temporal soil moisture patterns—A meta-analysis using plot to catchment scale data. *J. Hydrol.*, **520**, 326–341, doi:10.1016/j.jhydrol.2014.11.042.

Kristensen, K. J., and S. E. Jensen, 1975: A model for estimating actual evapotranspiration from potential evapotranspiration. *Nord. Hydrol.*, **6**, 170–188.

Larsen, M. A. D., J. H. Christensen, M. Drews, M. B. Butts, and J. C. Refsgaard, 2016a: Local control on precipitation in a fully coupled climate–hydrology model. *Sci. Rep.*, **6**, 22927, doi:10.1038/srep22927.

——, J. C. Refsgaard, K. H. Jensen, M. B. Butts, S. Stisen, and M. Mollerup, 2016b: Calibration of a distributed hydrol-

ogy and land surface model using energy flux measurements. *Agric. For. Meteor.*, **217**, 74–88, doi:10.1016/j.agrformet.2015.11.012.

Liang, S., C. J. Shuey, A. L. Russ, H. Fang, M. Chen, C. L. Walthall, C. S. T. Daughtry, and R. Hunt, 2003: Narrowband to broadband conversions of land surface albedo: II. Validation. *Remote Sens. Environ.*, **84**, 25–41, doi:10.1016/S0034-4257(02)00068-8.

Mariethoz, G., and B. F. J. Kelly, 2011: Modeling complex geological structures with elementary training images and transform-invariant distances. *Water Resour. Res.*, **47**, W07527, doi:10.1029/2011WR010412.

——, and J. Caers, 2014: *Multiple-Point Geostatistics: Stochastic Modeling with Training Images.* Wiley-Blackwell, 376 pp.

——, P. Renard, and J. Straubhaar, 2010: The direct sampling method to perform multiple-point geostatistical simulations. *Water Resour. Res.*, **46**, W11536, doi:10.1029/2008WR007621.

Mascaro, G., E. R. Vivoni, and L. A. Méndez-Barroso, 2015: Hyperresolution hydrologic modeling in a regional watershed and its interpretation using empirical orthogonal functions. *Adv. Water Resour.*, **83**, 190–206, doi:10.1016/j.advwatres.2015.05.023.

Mitchell, K. E., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.*, **109**, D07S90, doi:10.1029/2003JD003823.

Mittermaier, M., N. Roberts, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteor. Appl.*, **20**, 176–186, doi:10.1002/met.296.

Overgaard, J., 2005: Energy-based land-surface modelling: New opportunities in integrated hydrological modelling. Ph.D. thesis, Institute of Environment and Resources, Technical University of Denmark, 39 pp. [Available online at http://orbit.dtu.dk/files/6578332/MR2005-052.pdf.]

Pérez, C., G. Mariethoz, and J. M. Ortiz, 2014: Verifying the high-order consistency of training images with data for multiple-point geostatistics. *Comput. Geosci.*, **70**, 190–205, doi:10.1016/j.cageo.2014.06.001.

Perry, M. A., and J. D. Niemann, 2007: Analysis and estimation of soil moisture at the catchment scale using EOFs. *J. Hydrol.*, **334**, 388–404, doi:10.1016/j.jhydrol.2006.10.014.

——, and ——, 2008: Generation of soil moisture patterns at the catchment scale by EOF interpolation. *Hydrol. Earth Syst. Sci.*, **12**, 39–53, doi:10.5194/hess-12-39-2008.

Pokhrel, P., and H. V. Gupta, 2011: On the ability to infer spatial catchment variability using streamflow hydrographs. *Water Resour. Res.*, **47**, W08534, doi:10.1029/2010WR009873.

Qiu, J., X. Mo, S. Liu, and Z. Lin, 2014: Exploring spatiotemporal patterns and physical controls of soil moisture at various spatial scales. *Theor. Appl. Climatol.*, **118**, 159–171, doi:10.1007/s00704-013-1050-6.

Razavi, S., and H. V. Gupta, 2015: What do we mean by sensitivity analysis? The need for comprehensive characterization of "global" sensitivity in Earth and environmental systems models. *Water Resour. Res.*, **51**, 3070–3092, doi:10.1002/2014WR016527.

Renard, P., and D. Allard, 2013: Connectivity metrics for subsurface flow and transport. *Adv. Water Resour.*, **51**, 168–196, doi:10.1016/j.advwatres.2011.12.001.

——, J. Straubhaar, J. Caers, and G. Mariethoz, 2011: Conditioning facies simulations with connectivity data. *Math. Geosci.*, **43**, 879–903, doi:10.1007/s11004-011-9363-4.

Ringgaard, R., M. Herbst, T. Friborg, K. Schelde, A. G. Thomsen, and H. Soegaard, 2011: Energy fluxes above three disparate surfaces in a temperate mesoscale coastal catchment. *Vadose Zone J.*, **10**, 54–66, doi:10.2136/vzj2009.0181.

Roberts, N. M., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, **15**, 163–169, doi:10.1002/met.57.

——, and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Rosenbaum, U., H. R. Bogena, M. Herbst, J. A. Huisman, T. J. Peterson, A. Weuthen, A. W. Western, and H. Vereecken, 2012: Seasonal and event dynamics of spatial soil moisture patterns at the small catchment scale. *Water Resour. Res.*, **48**, W10544, doi:10.1029/2011WR011518.

Rouse, J. W., Jr., R. H. Haas, J. A. Schell, and D. W. Deering, 1973: Monitoring the vernal advancement and retrogradation (green wave effect) of natural vegetation. Progress Rep. RSC 1978-1, 112 pp. [Available online at https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19730017588.pdf.]

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

Savitzky, A., and M. J. E. Golay, 1964: Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, **36**, 1627–1639, doi:10.1021/ac60214a047.

Sheffield, J., and Coauthors, 2014: A drought monitoring and forecasting system for sub-Sahara African water resources and food security. *Bull. Amer. Meteor. Soc.*, **95**, 861–882, doi:10.1175/BAMS-D-12-00124.1.

Shrestha, P., M. Sulis, M. Masbou, S. Kollet, and C. Simmer, 2014: A scale-consistent terrestrial systems modeling platform based on COSMO, CLM, and ParFlow. *Mon. Wea. Rev.*, **142**, 3466–3483, doi:10.1175/MWR-D-14-00029.1.

Shuttleworth, W. J., and J. S. Wallace, 1985: Evaporation from sparse crops—An energy combination theory. *Quart. J. Roy. Meteor. Soc.*, **111**, 839–855, doi:10.1002/qj.49711146910.

Sobol', I. M., and S. Kucherenko, 2009: Derivative based global sensitivity measures and their link with global sensitivity indices. *Math. Comput. Simul.*, **79**, 3009–3017, doi:10.1016/j.matcom.2009.01.023.

Stisen, S., M. F. McCabe, J. C. Refsgaard, S. Lerer, and M. B. Butts, 2011: Model parameter analysis using remotely sensed pattern information in a multi-constraint framework. *J. Hydrol.*, **409**, 337–349, doi:10.1016/j.jhydrol.2011.08.030.

——, A. L. Højberg, L. Troldborg, J. C. Refsgaard, B. S. B. Christensen, M. Olsen, and H. J. Henriksen, 2012: On the importance of appropriate precipitation gauge catch correction for hydrological modelling at mid to high latitudes. *Hydrol. Earth Syst. Sci.*, **16**, 4157–4176, doi:10.5194/hess-16-4157-2012.

Tian, W., X. Li, G.-D. Cheng, X.-S. Wang, and B. X. Hu, 2012: Coupling a groundwater model with a land surface model to improve water and energy cycle simulation. *Hydrol. Earth Syst. Sci.*, **16**, 4707–4723, doi:10.5194/hess-16-4707-2012.

Vancutsem, C., P. Ceccato, T. Dinku, and S. J. Connor, 2010: Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa. *Remote Sens. Environ.*, **114**, 449–465, doi:10.1016/j.rse.2009.10.002.

Wanders, N., M. F. P. Bierkens, S. M. de Jong, A. de Roo, and D. Karssenberg, 2014: The benefits of using remotely sensed soil moisture in parameter identification of large-scale hydrological models. *Water Resour. Res.*, **50**, 6874–6891, doi:10.1002/2013WR014639.

Wealands, S. R., R. B. Grayson, and J. P. Walker, 2005: Quantitative comparison of spatial fields for hydrological model assessment—Some promising approaches. *Adv. Water Resour.*, **28**, 15–32, doi:10.1016/j.advwatres.2004.10.001.

Western, A. W., G. Blöschl, and R. B. Grayson, 1998: Geostatistical characterisation of soil moisture patterns in the Tarrawarra catchment. *J. Hydrol.*, **205**, 20–37, doi:10.1016/S0022-1694(97)00142-X.

——, ——, and ——, 2001: Toward capturing hydrologically significant connectivity in spatial patterns. *Water Resour. Res.*, **37**, 83–97, doi:10.1029/2000WR900241.

Williams, J. L., and R. M. Maxwell, 2011: Propagating subsurface uncertainty to the atmosphere using fully coupled stochastic simulations. *J. Hydrometeor.*, **12**, 690–701, doi:10.1175/2011JHM1363.1.

Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance, and B. G. Brown, 2014: Beyond the basics: Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Wea. Forecasting*, **29**, 1451–1472, doi:10.1175/WAF-D-13-00135.1.

Wood, E. F., and Coauthors, 2011: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resour. Res.*, **47**, W05301, doi:10.1029/2010WR010090.

Xia, Y., and Coauthors, 2016: Basin-scale assessment of the land surface water budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *J. Geophys. Res. Atmos.*, **121**, 2750–2779, doi:10.1002/2015JD023733.