AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR016607

Key Points:
- Quantifying the human perception in respect to pattern comparisons
- Performance metrics for spatial evaluation of distributed hydrological models
- Identifying complex spatial model defects by including new data in the model evaluation

Correspondence to:
J. Koch,
juko@geus.dk

Citation:
Koch, J., K. Hø. 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.

Received 28 OCT 2014
Accepted 22 JAN 2015
Accepted article online 31 JAN 2015
Published online 26 FEB 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

Julian Koch[1,2], Karsten Høgh Jensen[1], and Simon Stisen[2]

[1]Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark, [2]Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark

Abstract The hydrological modeling community is aware that the validation of distributed hydrological models has to move beyond aggregated performance measures, like hydrograph assessment by means of Nash-Suitcliffe efficiency toward a true spatial model validation. Remote sensing facilitates continuous data and can be measured on a similar spatial scale as the predictive scale of the hydrological model thereby it can serve as suitable data for the spatial validation. The human perception is often described as a very reliable and well-trained source for pattern comparison, which this study wants to exploit. A web-based survey that is interpreted based on approximately 200 replies reflects the consensus of the human perception on map comparisons of a reference map and 12 synthetic perturbations. The resulting similarity ranking can be used as a reference to benchmark various spatial performance metrics. This study promotes Fuzzy theory as a suitable approach because it considers uncertainties related to both location and value in the simulated map. Additionally, an EOF-analysis (Empirical Orthogonal Function) is conducted to decompose the map comparison into its similarities and dissimilarities. A modeling case study serves to further examine the metrics capability to assess the goodness of fit between simulated and observed land surface temperature maps. The EOF-analysis unambiguously identifies a systematic depth to groundwater table-related model deficiency. Kappa statistic extended by Fuzziness is a suitable and commonly applied measure for map comparison. However, its apparent bias sensitivity limits it's capability as a diagnostic tool to detect the distinct deficiency.

1. Introduction

Physically based distributed hydrological models provide the possibility to simulate the spatial distribution of hydraulic processes within the catchment as well as discharge estimates along the stream network [*Beven and Binley*, 1992; *Refsgaard*, 1997]. Due to their ability to make spatial explicit predictions, distributed models are often favored over lumped model approaches because they represent to some degree the spatial heterogeneity of the hydrological processes within the catchment. Factors like land use or climate that constrain the hydrological cycle are generally distributed in space. Hence predictive spatial modeling of hydrological responses is a logical and required framework which can build upon a vast availability of spatial data with a large toolbox of GIS systems, sophisticated model codes, and sufficient computational power [*Liu et al*., 2012]. Moreover, in order to optimize remediation and environmental policy-making, modeling applications are required to simulate not only the quantity and quality of water in the stream but also where contaminates originate from [*J. C. Refsgaard et al*., 2014]. Hydrologists have traditionally evaluated models using an efficiency measure describing the goodness of fit of the simulated discharge, like e.g., the Nash- Sutcliff [*Gupta et al*., 2009]. Discharge, just like hydraulic head, is an aggregated measure which represents all hydraulic processes that take place upstream of the measurement's location. The DMIP (Distributed Model Intercomparison Project) compares distributed and lumped models in respect to their ability to predict discharge at the watershed outlet and rates them as equally well performing [*Smith et al*., 2012]. However, the strength of distributed models over lumped models is their ability to simulate discharge additionally at interior catchment gauges [*Pokhrel and Gupta*, 2011] while successfully preserving the overall water balance. Even though interior stream gauges are simulated correctly [*Smith and Gupta*, 2012], it does not warrant the

conclusion that hydraulic processes at grid scale are simulated correctly. *Refsgaard* [2000] and *Beven and Feyen* [2002] suggested to formally move practices in distributed hydrological modeling toward adequate validation by considering continuous spatial observation data. Spatial data, like snow-cover data derived from remote sensing [*Warscher et al*., 2013], dense soil moisture networks [*Cornelissen et al*., 2014; *Sciuto and Diekkruger*, 2010], and land-surface temperature data [*Silvestro et al*., 2013; *Stisen et al*., 2011a] have already been incorporated into the model validation and assure that the validation is undertaken at a similar scale to the model's prediction. However, there exists no formal guideline on how to assess the goodness of fit of the spatial explicit model predictions. Spatial model evaluation is an active field of research in other disciplines besides hydrology, e.g., atmospheric sciences [*Brown et al*., 2011; *Gilleland et al*., 2010]. In more detail, the Model Evaluation Tool (MET) [*Brown et al*., 2009] offers a broad variety of tools for "object-based" and "neighborhood" evaluation techniques that allow a quantitative assessment of a spatially explicit forecast. Within hydrological science, *Grayson et al*. [2002] underlined the need for a true spatial evaluation of distributed model, which *Wealands et al*. [2005] reinforced by reviewing several spatial performance metrics. Their main conclusion was that simple global statistics that operate locally through cell-wise comparison of the simulated and observed maps of hydrological variables, like e.g., mean-error, root mean square error or correlation coefficient are not sufficient. These metrics are limited as they operate locally and do not consider information on patterns or the spatial correlation of the data. Along these lines, this study aims at extending the work of *Wealands et al*. [2005] by identifying suitable spatial performance metrics that go beyond standard local statistics, for the validation of a distributed hydrological model. Two approaches are implemented in that context, namely Fuzzy theory and Empirical Orthogonal Functions (EOF). Fuzziness [*Hagen*, 2009] captures the "vagueness" of a map and can be represented via uncertainty in location and value. The EOF-analysis is most commonly applied to understand predominant spatial or temporal patterns in observed data [*Graf et al*., 2014] by means of decomposition. In this study, the application will be extended in a novel approach to account for a quantitative comparison of simulated and observed data. *Hagen* [2003] and *Wealands et al*. [2005] regard the human perception as the most powerful and reliable source for comparing spatial patterns. Despite the subjectivity of individuals, a visual comparison of observed and simulated spatial data is reliable. The human perception can quickly integrate information on not only the overall similarity but also, the similarity of specific features and even spatial shifts or alterations of objects. This study exploits the human competence by incorporating a web-based survey of the human perception to compare and benchmark existing and novel spatial performance metrics. The survey contained 12 synthetically perturbed maps and the test persons were asked to evaluate their similarity in respect to a reference map. Additionally, a modeling case study is incorporated to test several spatial performance metrics on a physical distributed model (MIKE-SHE), where land-surface temperature (LST) data derived from the Moderate Resolution Image Radiospectrometer (MODIS) sensor is incorporated for model validation. Ultimately, a true spatial model evaluation that quantifies the spatial model performance in a geostatistical manner has to be implemented in multiobjective calibration frameworks of distributed hydrological models [*Efstratiadis and Koutsoyiannis*, 2010; *Pokhrel et al*., 2012]. However, further tests about stability and robustness must be conducted before these metrics are fit for an inverse calibration. Until then they can serve as diagnostic tools to learn about distributed models and to identify possible spatial model deficiencies that would normally be overlooked by ordinary model evaluation, through e.g., hydrograph assessment.

The overall goals of this study are (1) to develop methods for spatial evaluation of distributed hydrological models, (2) to benchmark spatial performance metrics against a web-based survey of the human perception, and (3) to translate the insights gained from the survey to a real modeling-case study.

2. Study Area

The study area is the 1050 km^2 large Ahlergaarde catchment, which is a subcatchment to the Skjern river (2500 km^2) in western Jutland, Denmark. The Skjern river outlet is the Ringkøbing Fjord and all other boundary conditions are defined by topographic divides. Topography shows gentle elevation gradients toward west with a maximum of 125 m.a.s.l.. Land use is predominately characterized by agriculture and coniferous forest. The catchment's climate can be described as temperate maritime climate, with mean annual precipitation of 990 mm and mean annual temperature of 8.2°. Since 2007, the Skjern catchment has constituted the Danish Hydrological Observatory, HOBE [*Jensen and Illangasekare*, 2011], and has been subject to

numerous experiments and measurements concerning precipitation, evapotranspiration, greenhouse gas exchange, ground-surface water interactions, and other related topics. This makes it highly suitable for an intensive model evaluation/calibration study with focus on spatial model evaluation.

3. Hydrological Model and Validation Data

This study builds upon the integrated modeling activities in the HOBE project. The existing model is based on a unique large data set and was subject to a thorough calibration. *Stisen et al*. [2011b] present the complete derivation of climate forcing data used for the current model setup based on 16 climate stations inside and around the catchment. The detailed implementation of the current model setup in the Mike-SHE SW-ET model is given by *Stisen et al*. [2011a]. SW-ET is a two source energy balance land surface model based on the *Shuttleworth and Wallace* [1985] model and enhances the physical basis of the distributed Mike-SHE model [*Abbott et al*., 1986]. The distributed coupling of surface, subsurface, and atmosphere builds a powerful predictive tool for the assessment of surface processes that are spatially explicit. The subsurface component in the model constrains a pronounced energy exchange with the atmosphere at locations with a shallow groundwater table and thus large water availability. Land surface temperature (LST) is a central variable in the SW-ET model [*Overgaard*, 2005] and can be captured via remote sensing (MYD11A1 Aqua-MODIS) and then be utilized for model validation. The model runs on hourly time steps and therefore the validation against instantaneous MODIS observation is very meaningful. LST data from the MODIS satellite are available throughout the entire year. However, this study considers only days with minimum 95% cloud-free pixels per image for model validation. This yields in total 33 maps in the summer months from April to September for a 6 year simulation period (2007–2012). The LST data, besides other hydrological variables like discharge, actual ET, soil moisture, and hydraulic head measurements are included in the previous model calibration by *Stisen et al*. [2011a]. The design and results of this multiconstrained calibration exercise of the HOBE research catchment are subject of a separate paper. The elaborated, physically based, and distributed hydrological model of Skjern river in combination with the LST data set is recently applied by *Guzinski et al*. [2014] to compare different evapotranspiration models.

4. Methods

4.1. Survey

4.1.1. Human Perception

One crucial function of human vision is perception. This allows for a person to make visual judgments that are supported by their cognitive system, e.g., memory, semantics, spatial reasoning, planning and communication, and hence make them extremely reliable [*Jacob and Jeannerod*, 2003]. In general, human perception can be understood as both a process and a product. The latter case has been studied intensively by epistemologists who attest perception is of paramount importance because it creates knowledge [*Carterette and Friedman*, 1974]. The process of perception can be analyzed from a neurophysiological viewpoint [*Spillmann and Werner*, 1990] where the connection between eye and brain is investigated as neural mechanism to explain how humans perceive. Gestalt psychology by *Wertheimer* [1912] studies human vision and characterizes it by perceptual grouping; grouping through proximity, similarity, good continuation, etc. These grouping principles are important when discussing how humans perceive, for example, spatial patterns. The theory is well presented in a centennial anniversary review by *Wagemans et al*. [2012].

4.1.2. Survey Idea

As introduced above, the human perception in combination with a person's cognitive system, builds a reliable source for pattern recognition and map comparison. The human perception allows to quickly select an appropriate scale to investigate a spatial pattern as well as to identify the predominant underlying spatial structures of a map. This is very applicable for a credible pattern comparison because the human perception will easily recognize similarities and dissimilarities between two maps. However, it is difficult to establish a quantification of human judgment and, thus, to determine exactly how similar/dissimilar two maps are. Also, reliable visual judgment cannot be granted if hundreds of maps were to be compared entirely qualitatively, which is often the case for model optimization tasks. Therefore, quantitative algorithms (spatial performance metrics) for pattern comparison will always be the obvious choice for a model calibration. Nevertheless, in order to retain some of the advantages of qualitative map comparisons, this study features

a web-based survey with the aim to utilize the results as a reference to test and tune various spatial performance metrics and to benchmark their performance. A similar study, where a web-based survey is conducted to incorporate the human perception into quantitative map comparison is published by *Kuhnert et al*. [2005] with the focus on spatial modeling in landscape ecology. The survey results were successfully used to calibrate spatial metrics and to benchmark them. The benchmark test revealed that the Fuzzy Kappa index and a moving window map comparison approach prove best with a strong correlation of 0.844 and 0.845, respectively. In a second study by *Fritz and Lee* [2005], a different approach to quantify human perception was applied. Expert exaltation was incorporated to define the relation between land use categories for a Fuzzy Kappa assessment.

4.1.3. Survey Design

One LST-map derived from a MODIS LST-scene serves as the reference map in the survey. It is chosen because its mean, standard deviation, and variogram best represent a "normal" LST-pattern within the variability of all 33 LST-maps available. In total, 12 different synthetic perturbation strategies are imposed on the reference map in order to compile the data set for the survey (Figure 1). The perturbation strategies are given in Table 1 and are chosen to reflect possible model errors, e.g., systematic bias, random deviations, or spatial shift.

4.1.4. Survey Interpretation

Direct quantification of human judgment of a map comparison is difficult. Therefore, the test persons are simply asked to decide which one of two given perturbations is more similar or if they are equally similar in respect to the reference map. Comparing all 12 perturbations with each other yields 66 ((12^2–12)/2) comparisons. By eliminating obvious comparisons beforehand, e.g., ID #1 and ID #7 in Figure 1, the number of comparisons are reduced to 48 in order to make the survey more compact. The web-based questionnaire tool "KwikSurveys" is used to conduct the survey and it will stay online for the time being (http://kwiksurveys.com/s.asp?sid=05a50l3t1jb4eyp326466).

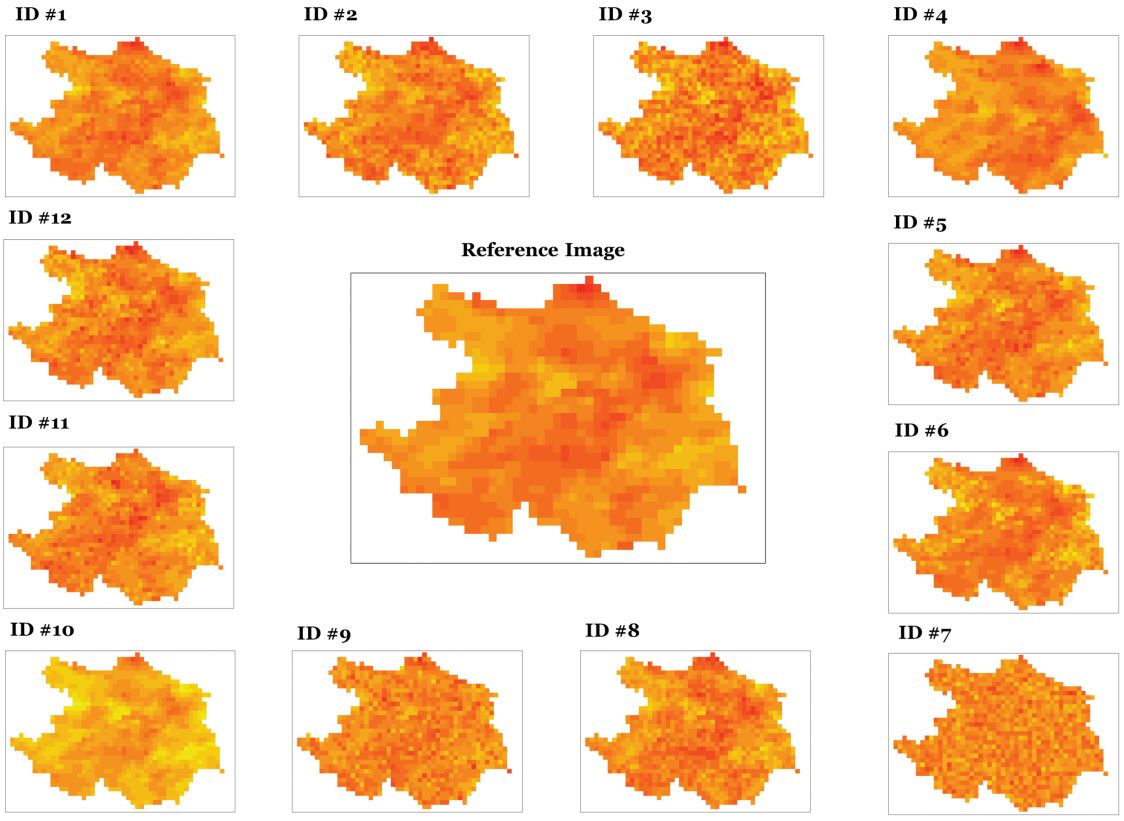

Figure 1. The 12 perturbations, following Table 4, that are used in the survey. All synthetic changes are imposed on the reference map in the center. The maps depict LST, where darker colors represent warmer temperature. All maps underlie the same symbology; minimum and maximum values can be neglected at this point.

Table 1. The Perturbation Strategies Used to Generate the Synthetic Maps for the Survey

ID	Description
1	30% probability of +1/0/−1 change in category.
2	Probability of +2/+1/0/-1/-2 change in category decreases towards the center.
3	100% probability of +2/+1/0/-1/-2 change in category.
4	Shift of +5 cells in X and Y direction. Reference is top left.
5	30% probability of +2/+1/0/-1/-2 change in category.
6	70% probability of +2/+1/0/-1/-2 change in category is imposed to low temperature cells (<23deg.).
7	Cells are randomly assigned a category in respect to the distribution if the reference map.
8	100% Probability of +1/0/-1 Change in Category.
9	Cells are randomly assigned a category in respect to the distribution if the reference image. 50% of the cells are "conditioned" to the reference map.
10	A general bias of −2 degree is imposed to all cells.
11	Shift of −2 cells in X and Y direction. Reference is top left. Additionally, 30% probability of +2/+1/0/-1/-2 change in category
12	Probability of +2/+1/0/-1/-2 change in category decreased towards the lower right.

In addition to the 48 map comparisons, the test person's experience in working with spatial data is investigated because the survey is sent out to both science and private network. The methodology underlying the interpretation of the survey is original and novel in this study. All replies are compiled in a matrix containing all possible map comparisons (12x12): For every positive similarity reply, the map receives +1 and for every equal similarity reply +0.5 to the corresponding cell in the matrix, which is then divided by the total number of replies for the corresponding comparison. Comparisons that are eliminated beforehand to reduce the quantity of questions are reimbursed by 1 or 0.75, depending on how obvious the comparison is; e.g., ID#1 to ID#7 =1 and ID#1 to ID#8 =0.75 both in favor of the ID#1. The row sum for each map divided by 11 (number of comparison for each map) yields a similarity score between zero and one that can be used to rank the images according to the consensus of the human perception, where a score of one means full similarity and a score of zero full dissimilarity.

4.2. EOF

Empirical Orthogonal Functions (EOF) are a broadly used method for data analysis and exploration of continuous variables and have been used by *Perry and Niemann* [2007] and *Korres et al*. [2010] for the analysis of soil moisture patterns at catchment scale and by *Sun et al*. [2012] and *Ahmad et al*. [2014] for large-scale seasonal precipitation patterns. It is analogous to the principal component analysis but only focuses on one single variable and decomposes a space-time dataset into a set of orthogonal spatial patterns (EOFs) and a series of factors. The latter determines the importance of each EOF to the variance at each time step and is often referred to as loadings or expansion coefficients. This study focuses on the spatial variability of LST. The analysis starts with a matrix A of the space-time temperature data where a_{ij} denotes the temperature value at time *j* (columns) and position *i* (rows). This orientation of the data matrix focuses on spatial variability, whereas the matrix transpose of A would give the possibility to analyze its temporal variability. The matrix Z of temperature anomalies is derived by a prior mean removal:

$$z_{ij} = a_{ij} - \overline{a_j} \tag{1}$$

where z_{ij} is the temperature anomaly at position *i* and time *j* and $\bar{a}_j$ is the spatially averaged temperature at time *j*. After computing the covariance of matrix *Z*, the covariance matrix is diagonalized which yields its eigenvalues *E* and orthogonal eigenvectors *F*. Multiplication of the latter with *Z* yields the EOFs, where the first EOF is oriented in the direction of maximum variance, the subsequent EOFs are constrained to be perpendicular to the one before, and consecutively explain more of the total variance. The eigenvalues define the amount of contribution of the direction of each eigenvector. Therefore, the temperature anomalies can be represented by:

$$Z = F \times E^T \tag{2}$$

where *F* contains the EOFs as columns and E^T comprises the loadings as rows, note that the superscript T indicates the matrix transpose. The first EOF, which mirrors as much as possible of the variance of Z, can be used in combination with the first row of E^T as a noise-reduced memory map as it filters the most prevalent similarities in patterns. Any further EOF will consecutively add to the total variance of A. Further examples

of successful implementation of the EOF analysis in a soil moisture context are *Jawson and Niemann* [2007] and *Graf et al*. [2014], in atmospheric science [*Hannachi et al*., 2007] or in remote sensing applications [*Müller et al*., 2014]. These studies show that EOF analysis is a powerful tool for exploration of both observed and modeled data. Nevertheless, very few attempts have been made to extend the usual application of EOF-analysis to spatial model validation. *Wang et al*. [2006] evaluated simulated and observed rainfall patterns using EOF's; however, the evaluation was purely qualitative by visually comparing the first and the second EOF based on the simulated and observed precipitation data.

For the context of this study a quantification, thus a single robust performance metric is of interest. We tested to conduct the EOF analysis individually on both the simulated and observed space-time matrix of temperature: 1021 rows representing the cells and 33 columns representing the days. However, due to differences in mean temperature at each day, a comparison of eigenvectors and/or eigenvalues is flawed. Also, z-transforming the matrices to have a mean of zero and a standard deviation of one affects the comparison because some of the internal variability is dampened. *Graf et al*. [2014] successfully analyzed a soil moisture data set by decomposition of the variance into eigenvectors and their eigenvalues. Subsequently, a k-means cluster analysis of the loadings of the first two EOFs was conducted to filter the two most prominent underlying patterns of soil moisture. A similar procedure is conceivable for model evaluation, although the robustness of the k-means clustering is not granted in an automatic calibration exercise as unrealistic clusters might occur. Instead, we chose to build one integral data matrix containing both, observed and simulated data: 1021 rows and 66 columns. The sequence of rows and columns in the matrix is of no importance in an EOF analysis. The combined EOF analysis yields orthogonal EOF maps that explain the combined intervariability and intravariability of both data sets. It is anticipated that after decomposition each EOF represents a particular similarity or dissimilarity between the observed and simulated data sets. The loadings express how much each day, simulated or observed, contributes to the direction of the corresponding EOF. Assuming that a LST map is simulated in perfect agreement suggests that the loadings for dayX_{sim} and day-X_{obs} are essentially equal. Consequently, the loading deviation can be understood as an indicator for similarity; low deviation indicates high similarity. The variance contribution decreases consecutively for the EOFs, hence the loading deviation of EOF1 is supposed to contribute most significantly. Thus the loading deviation needs to be weighted to represent this accordingly. The variance contribution of each EOF does represent both the intravariability of the observed and simulated data, but also the variability between those and can be used to weight the respective loading deviation. The EOF similarity score between a simulated and observed pattern at day x can be formulated as follows:

$$S_{EOF}^{x} = \sum_{i=1}^{n} \left| w_i \left(load_i^{simx} - load_i^{obsx} \right) \right| \tag{3}$$

where w_i, the variance contribution of the *i*'th EOF, is multiplied by the deviation between the corresponding EOF's loadings for simulated and observed day x.

4.3. Kappa Statistics

Kappa statistics are widely applied methods for map comparison in distributed environmental modeling [*Bennett et al*., 2013] aiming at e.g., spatial model validation [*Rose et al*., 2009; *Sciuto and Diekkruger*, 2010; *van Vliet et al*., 2013] or the analysis of land-use change dynamics [*Hagen and Martens*, 2008; *Pontius et al*., 2004; *Power et al*., 2001]. It is an overall measure for similarity between two categorized maps and follows the original idea by *Cohen* [1960] who assessed nominal scales. The comparison of numerical maps is feasible. However, they have to be categorized for proper application. In this study, a categorization into one degree intervals is performed prior to the pattern evaluation. The general Kappa index uses the actual observed agreement (P_o) between two maps and "corrects" it for the expected agreement (P_e) through random allocation of categories following their histograms (equation (4)). The agreements are derived from a confusion matrix (Table 2).

$$K = \frac{P_o - P_e}{1 - P_e} \tag{4}$$

The outcome of the correction is a Kappa value ranging from 1 = perfect agreement, to 0 = expected agreement between two uncorrelated maps, and below 0 = the observed agreement is worse than the expected agreement.

Table 2. Virtual Confusion Matrix Used for Comparing a Simulated Map With a Reference Map[a]

		Reference Map				
Simulated Map	Category	1	2	...	N	Total
	1	P_{11}	P_{12}	...	P_{1N}	X_1
	2	P_{21}	P_{22}	...	P_{2N}	X_2
	...	...	...	...	...	...
	M	P_{M1}	P_{M2}	...	P_{MN}	X_M
	Total	Y_1	Y_2	...	Y_N	XY

[a] X_m represents the sum for each row; Y_n the sum for each column; XY the sum of all cells, and P_{mn} the total of all membership vectors modeled as category m and observed as category n in the reference map.

4.3.1. Fuzziness

Hagen [2003] extended the Kappa index by Fuzzy Set theory [*Zadeh*, 1965, 1968) to the Fuzzy Kappa index (FK) which allows for two distinct sources of uncertainty/vagueness in the map comparison. The two components of fuzziness which are considered are fuzziness of location (FoL) and fuzziness of category (FoC); tolerance in position, and value, respectively. The latter was already acknowledged by *Cohen* [1968] who first defined the weighted Kappa index. It implies that a fuzzy agreement is assigned for categories which are similar, thus with values close to each other for numerical maps, or categories that are logically related for nominal maps. FoL represents proximity relations between neighboring cells, thus similarity is regarded as a function of distance and can be described by a decay function:

$$Y(x) = e^{-\lambda x} \tag{5}$$

where x is the distance and ë a constant. In combination, the fact that some pairs of categories are more similar than others and taking proximity coherence into account builds a strong measure for overall similarity [*Hagen et al*., 2005]. This fuzzy representation yields a fuzzy membership vector that contains a cell's similarity to all other apparent categories.

4.3.2. Fuzzy Weighted Kappa

The FK approach by *Hagen* [2003] compares two independent maps by the so-called two-way comparison where the two membership vectors of a cell, one from each map, are treated equally. This study evaluates a simulated map against a reference map and thus it is preferred to simply compare the membership vector of the simulated cell to the category of the reference cell, which can be represented by a so-called crisp vector. This approach is implemented in this study and was coined by *Huang and Lees* [2007] as the fuzzy weighted Kappa and will from now one be referred to as Fuzzy Kappa (FK). The basis of this approach is a confusion matrix, also referred to as an error or coincide matrix [*Foody*, 2008; *Remmel*, 2009] which is converted into a virtual confusion matrix, by considering FoL. Without the consideration of FoL, the basic confusion matrix is a simple cell-by-cell comparison disregarding proximity relations. FoL contributes to the membership vector at each cell, by taking the crisp vector of the neighborhood cells according to equation (5) into consideration. The virtual confusion matrix is then built by summing all membership vectors up and it denotes as follows:

The observed agreement (P_o) is derived from the virtual confusion matrix by:

$$P_o = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} W_{mn}P_{mn}}{XY} \tag{6}$$

where W_{mn} is a weight matrix that reflects the FoC between category m and n. The diagonal values in W_{mn} will be equal to 1 in most cases; the remaining entries can be reflected by the distance decay function in equation (5), where distance is regarded as numeric deviation between the categories. The expected agreement (P_e) takes into account the histograms of the two maps:

$$P_e = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} W_{ij}X_mY_n}{XY^2} \tag{7}$$

where X_m and Y_n are the row and column sums of the virtual confusion matrix. *Hagen et al*. [2005] advocate correcting the observed agreement P_o by the expected agreement P_e because maps with unevenly distributed histograms bias P_0. Two main points of criticism in regard to the expected agreement as a baseline in the FK derivation (equation (4)) have been brought forward. First, *Pontius and Millones* [2011] argue using the expected agreement as a baseline and conclude to abandon Kappa as a measure for overall similarity because the expected agreement due to randomness completely disregards the quantity of local

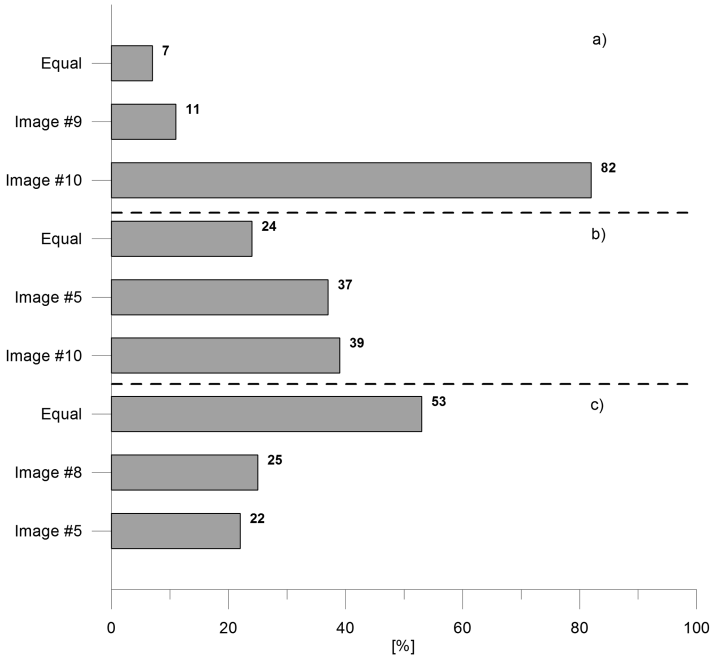

Figure 2. Three examples showing survey results: e.g., (a) compares the similarity between ID#9 and ID#10 (Figure 1); 7% rated them as equally similar to the reference, 11% attested ID#9 a higher similarity, and the remaining 82% rated ID#19 with a higher similarity to the reference map than ID#9. 195, 187, and 187 test persons answered question (a), (b) and (c), respectively.

disagreement. This discussion is, however, within the field of remote sensing. Second, *Hagen* [2009] present an improved Fuzzy Kappa statistic where the expected agreement accounts for spatial autocorrelation between two maps. It is uttered that spatial autocorrelation lowers the expected agreement and neglecting this effect will bias FK to be negative, although maps appear to have a considerable agreement.

4.4. Fuzzy Similarity Map

In order to obtain a spatial and gradual analysis of the similarity of two maps, a fuzzy agreement map is a better choice as it shows the similarity of each individual cell; FK, on the other hand, will only give an overall measure of similarity between two maps [*Wealands et al*., 2005]. Following *Hagen* [2003], the simulated map can be utilized to build crisp vectors for each cell, which do not involve fuzziness at all. The crisp vectors can then easily be transformed to fuzzy category vectors by adding information on category membership weights (FoC). The resulting fuzzy category vectors describe the similarity of each simulated cell to all other categories. FoL has to be incorporated in order to consider proximity relations to neighboring cells. This is achieved by scanning the neighborhood cells for each category and by accounting for a decreasing membership with distance following a decay function (equation (5)). *Hagen* [2003] presents a detailed example of how to proceed in order to derive the so-called fuzzy neighborhood vector at each cell, containing FoL and FoC. As opposed to *Hagen* [2003], this study utilizes a more simplified way to interpret the fuzzy neighborhood vector by simply multiplying it with the crisp vector of the coinciding cell in the reference map. This step yields a similarity score for each cell, considering fuzziness only for the simulated map and not for the reference map. The final Fuzzy Similarity (FS) map provides insight into spatial model deficiencies.

5. Results

The first part of the result section covers the survey interpretation and how the derived ranking of the 12 perturbed maps (Figure 1) is applied to benchmark spatial metrics and to test them against the human perception. In the second part, the given metrics are applied in a modeling case study, where the spatial performance is assessed by means of 33 LST map comparison.

5.1. Survey

5.1.1. Survey Interpretation

The number of replies varies between 188 and 246 per question, which seems to be a reasonable sample size to reflect the consensus of the human perception. As a comparison, *Kuhnert et al*. [2005] received 186 replies in their survey and successfully utilized them to benchmark various pattern comparison algorithms. 43% of the test persons claimed to have previous experience working with spatial data. However, no significant differences on how the two groups rated the similarity between the perturbations and the reference map are discovered and thus the test persons are considered as a unity in the further analysis. Figure 2

exemplifies the results of three map comparisons which reflect the variety of reply patterns: (a) shows an agreement on similarity, (b) a disagreement of similarity, and (c) an agreement on equality. Case (b) and (c) manifest similar in the resulting matrix and are not further distinguished in the analysis. Table 3 presents the survey ranking, where the numbers denote the IDs of the perturbed maps in Figure 1. The survey ranking is compared with rankings derived by various spatial performance metrics. If the perturbed image receives the same rank by the performance metric and by the survey, the ID is marked in bolded font. Additionally, Table 3 gives the correlation of each method to the survey results. Following the procedure explained in section 5.1.4, each perturbed map receives a numeric similarity score between 0 and 1, which allows a correlation analysis with other numeric performance metrics. The root mean squared error (RMSE), mean error (ME), and correlation coefficient (R) are known metrics and there exists a common understanding of their strengths, weaknesses, and peculiarities. The RMSE performs well in regard to the absolute ranking and correlation. In contrast, the ME attests a very poor reproduction of the survey ranking where only the high similarity maps are ranked accordingly. R shows the strongest correlation, however absolute ranks are mostly missed.

5.1.2. EOF Versus Survey

The calculation of the EOF similarity score (equation (3)) is illustrated in Figure 3 for the 12 perturbed LST maps incorporated in the survey. The EOF analysis is based on 1021 rows (one for each cell) and 13 columns (one reference map and 12 perturbations). The assumption is that the closer the loadings of a perturbation are to the loadings of the reference map, the higher the similarity is. The first EOF constitutes 60.5% of the total variance and the second EOF explains additional 7%. The similarity score in equation (3) considers all loading deviations according to the variance contribution of the corresponding EOF, thus EOF1 supplies the largest fraction. The plot in Figure 3 underlines that most of the perturbed maps agglomerate close to the reference map, underlining their high similarity. ID#9 is ranked low by the EOF similarity score because the difference in the loadings of EOF1 (delta EOF1) between the reference map and ID#9 clearly dominates, despite a small "delta EOF2." The largest "delta EOF1" is attested to ID#7, which is consequently ranked lowest in Table 3. The second EOF is most apparent in ID#4 and ID#11, which are the only maps perturbed by a spatial shift. Additionally, ID#3 which is subject to continuous perturbation at all cells shows a large deviation in EOF2 as well. The loadings of ID#10 (bias of −2deg) coincide with the reference map, which states a perfect agreement between the two maps and underlines the bias insensitivity of the EOF-analysis, due to the prior mean correction of the data set.

5.1.3. FK Versus Survey

The major advantage of applying fuzziness in a pattern comparison is that the user can define the membership values through an e.g., distance decay function (equation (5)). On the other hand, this notion might complicate the application because of the distinct subjectivity that lies in the definition of FoL and FoC. Therefore, this study tests four different FK scenarios (Table 4) and benchmarks them by the survey ranking. Each scenario is equipped with different λ values, controlling the distance decay function which defines the

Table 3. Resulting Survey Ranking (Top Represents High Similarity Between Perturbed Map and Reference Map) Compared to Various Metrics[a]

Survey	RMSE	ME	R	EOF	FK Initial	FK Survey	FK CAT	FK LOC	P_o Survey
1	**1**	**1**	10	10	**1**	**1**	**1**	**1**	**1**
6	5	8	1	1	5	5	5	5	5
8	**8**	6	5	5	12	6	6	6	6
12	6	**12**	6	6	6	**12**	**12**	**12**	**12**
5	12	4	8	8	8	8	2	11	9
2	**2**	9	12	12	**2**	**2**	8	8	8
10	9	5	2	2	9	9	3	2	2
11	3	7	3	3	**11**	**11**	9	9	**11**
4	11	3	9	9	3	**4**	11	**4**	**4**
3	4	2	11	11	4	**3**	10	**3**	7
9	7	11	4	4	7	7	4	7	3
7	10	10	**7**	**7**	10	10	**7**	10	10
Correlation	0.76	0.08	0.83	0.80	0.70	0.65	0.83	0.58	0.43

[a]Root-mean-squared-error (RMSE), mean-error (ME), correlation-coefficient (R), Fuzzy Kappa with initial-, calibrated-, focus on FoC- and focus on FoL-λ parameters (FK initial, FK survey, FK CAT and FK LOC). The observed agreement P_o is based on FK survey λ parameters. Bolded letters indicate a correct positioning of the rank.

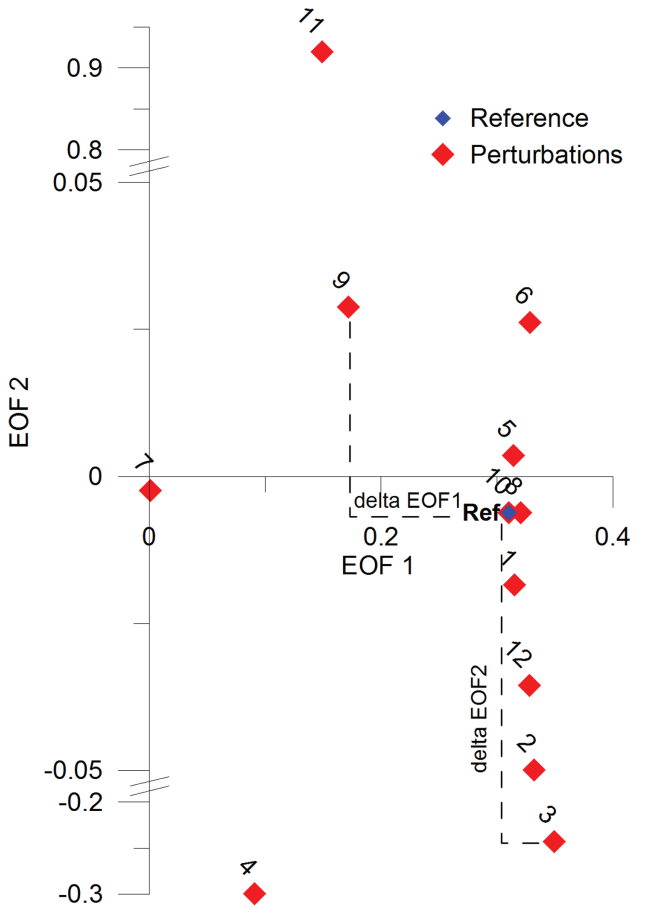

Figure 3. EOF-analysis for map comparison based on the survey maps (12 perturbations + reference). The axes reflect the unitless loadings of EOF1 and EOF2. "delta EOF1" and "delta EOF2" are elements of equation (3) and utilized to compute the overall EOF similarity score.

membership values for FoL and FoC. The 5 km radius for the FoL similarity space follows the average correlation length of the 33 observed LST maps. The FoC similarity space considers a deviation of maximum five categories. This threshold is unreasoned; however, it is tested and appears to be less sensitive than λ_{FoC}. The "FK initial" scenario, which is an initial qualified guess of the λ values puts focus on FoL (Table 4). λ_{FoL} and λ_{FoC} are calibrated in the "FK survey" scenario with the target to have as many absolute ranking matches with the survey as possible. The calibration is conducted in a semiautomatic manner with multiple λ_{FoL} and λ_{FoC} combinations with values varying between 0.1 and 2.5. The best result, with six direct rank matches, is obtained with 0.8 and 1, respectively (Table 4). Two further scenarios, namely "FK CAT" and "FK LOC" are incorporated which are almost entirely driven by FoC and FoL, respectively (Table 4). Concentrating on FoC and FoL individually can help to understand the implications of each individual fuzzy source. Additionally, the observed agreement P_o from the FK equation (equation (4)) with the same fuzzy parameters as "FK survey" is included in Table 3. The "FK survey" scenario performs clearly best at reproducing the ranking of the survey, although it shows a rather weak correlation of 0.65. The observed agreement (P_o) indicates a weak correlation to the survey ranking but is supplemented with an overall good positioning of the ranks. The qualified guess of λ_{FoL} and λ_{FoC} in the "FK initial" scenario performs well and overall similar to the "FK survey" scenario in reproducing the absolute ranking of the survey. "FK LOC" favors perturbations ID#4 and ID #11 compared to the other scenarios, which both are altered by a spatial shift. This is in good agreement to the pronounced FoL in the "FK LOC" scenario, which especially focuses on proximity relations of neighboring cells. Perturbation ID#10, a general bias of -2° is rated extremely different by the five methods. RMSE and ME are obviously very sensitive to a large bias, on the contrary R and EOF place ID#10 on the first place as both are corrected by the mean and thus are bias insensitive. The elevated rank of ID#10 in the "FK CAT" scenario, compared to the other FK scenarios, underlines that the bias sensitivity of the FK method can be controlled by λ_{FoC}. All metrics estimate the similarity score of ID#9 higher than the survey. In this case, the human perception is tricked because 50% of the cells in ID#9 are conditioned to the reference map and the other half is perturbed by noise. This perturbation is favored by all algorithms as 50% of the cells are identical, which yields a high similarity score. However, the scattered map cheats the human perception and thus the survey attests a low similarity score to ID#9.

5.2. Modeling Study

The spatial validation of the modeling case study of the Ahlergaarde catchment consists of 33 LST maps derived from the MODIS satellite that evaluate the spatial performance of the distributed Mike-SHE SW-ET model. A set of spatial performance metrics is applied for this exercise. EOF-analysis with focus on the

Table 4. Similarity Space for Fuzziness of Location (FoL) and Fuzziness of Category (FoC) Computed by Equation (5) for Different Scenarios: Initial Guess, Calibrated After Survey Ranking, Focus on FoL and Focus on FoC Reflect Columns (1) to (4), Respectively

		Similarity			
FoL		Initial	Survey	LOC	CAT
Ring	d (km)	$\lambda = 1$	$\lambda = 0.8$	$\lambda = 2$	$\lambda = 0.1$
0	0	1	1	1	1
1	1	0.37	0.45	0.14	0.90
2	1.4	0.24	0.32	0.06	0.87
3	2	0.14	0.20	0.02	0.82
4	2.2	0.11	0.17	0.01	0.80
5	2.8	0.06	0.11	0	0.75
6	3	0.05	0.09	0	0.74
7	3.2	0.04	0.08	0	0.73
8	3.6	0.03	0.06	0	0.70
9	4	0.02	0.04	0	0.67
10	4.1	0.02	0.04	0	0.66
11	4.2	0.01	0.03	0	0.65
12	4.5	0.01	0.03	0	0.64
13	5	0.01	0.02	0	0.61
FoC d (cat)		$\lambda = 1$	$\lambda = 0.8$	$\lambda = 2$	$\lambda = 0.1$
0		1	1	1	1
1		0.61	0.37	0.90	0.14
2		0.37	0.14	0.82	0.02
3		0.22	0.05	0.74	0
4		0.14	0.02	0.67	0
5		0.08	0.01	0.61	0

loadings (equation (3)) to quantify pattern similarity between two maps is as such a novelty in spatial model evaluation. Therefore, one section is dedicated to better understand the outputs of the EOF-analysis and its implications. Further the set of performance metrics is applied on the time series of 33 LST map comparisons to present differences and peculiarities of the individual metrics. At last, it is tested if the results of the spatial model validation can serve as a diagnostic tool to detect systematic model deficiencies.

5.2.1. Interpretation of EOF

This study promotes the use of the EOF-analysis as a tool to assess spatial model performance. It is anticipated that the decomposition of the space time LST data set of both MODIS data and model output yields in a clear distinction between similarities and dissimilarities between the two data sets. First the loadings of the simulated and observed maps are compared in Figure 4 and second the physical relevancy of the EOF maps is investigated in Figure 5. Figure 4 compares the observed and simulated loadings of the first two EOFs for all 33 LST maps. A clear distinction can be noticed between the first EOF that explains 49% of the total variance and the second EOF that calls for additional 18.5%, in combination they explain 67.5%. It appears that the first EOF captures predominantly similarities between the observed and simulated data set as the loadings manifest a strong correlation. Thus EOF1 highlights the predominant temperature pattern that is found in both, the observed and the simulated LST maps. The loadings of EOF2 are very scattered which implies that mainly dissimilarities are covered by the second EOF. In general, the closer the data plot to the diagonal in Figure 4, the higher the similarity is between the two maps. In order to further understand the physical relevancy of the EOFs, the first two EOFs are plotted against average temperature values for the 33 observed and simulated LST maps (Figure 5). EOF1 shows a strong positive correlation for both, simulated and observed average temperature. On the contrary, the correlation between averaged temperature and EOF2 depicts a very different picture. The observed data set still shows a strong positive

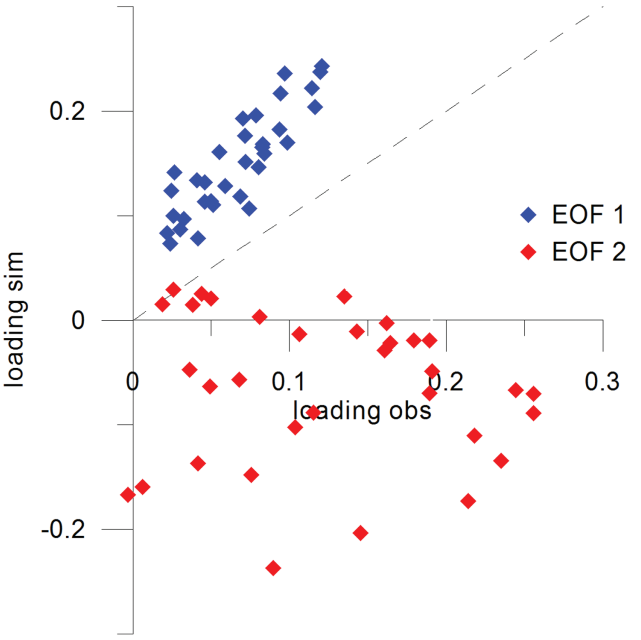

Figure 4. The unitless loadings for EOF1 and EOF2 for the 33 observed and simulated LST maps. The closer a point is to the dashed diagonal the higher is the similarity.

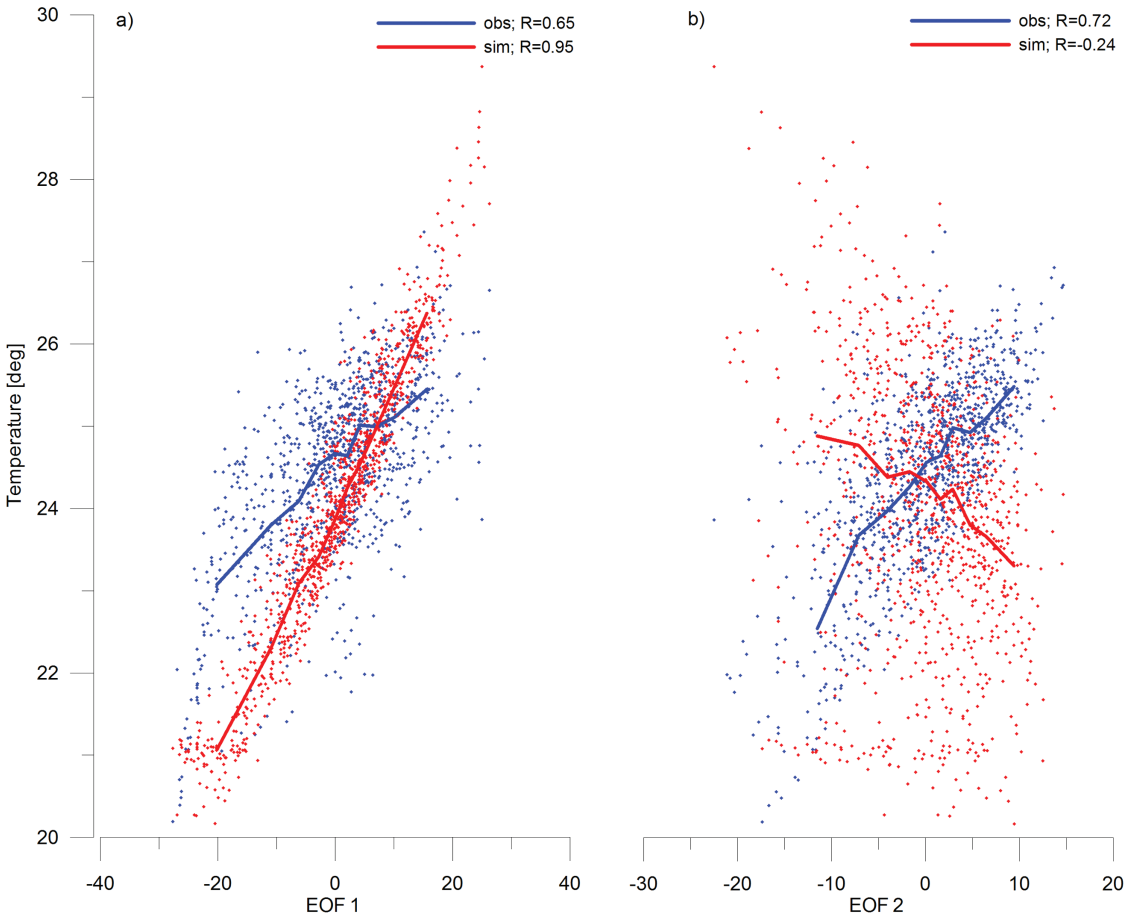

Figure 5. The scatterplots represent the maps of (a) EOF1 and (b) EOF2 against the average simulated and observed temperature based on all 33 LST maps. The trend curves are based on grouping the data on the x axis into 10 equal groups and plotting the median of each group on both axes.

correlation to the EOF2 map, opposed to the scattered simulated mean temperature that promotes a weak correlation to EOF2. Moreover this supports the conclusion from Figure 4 that the first EOF map filters the similarities between the two data sets. From Figures 4 and 5, it can also be concluded that the underlying temperature patterns of the observed data set are more persistent than the simulated patterns. This is strengthened by the disperse temperature correlation of simulated LST maps and EOF2 in Figure 5 and the loadings plot of both EOFs in Figure 4. The loadings for the observed data set in Figure 4 are almost entirely positive, whereas the simulated LST maps are associated with both, negative and positive loadings for EOF2. The change in loading sign indicates that the contribution of EOF2 is inverted at certain days, which makes the simulated LST maps less persistent than the observed data set. The 49% of variance that is explained by the first EOF can be interpreted as the error reduced similarity between the simulated and the observed LST maps and represents the basic pattern of high and low temperature cells that is represented in both data sets. Hypothetically, the subsequent EOFs are therefore a representation of the dissimilarities, with EOF2 being the most prominent, with 18.5% of the total variance. This study does only investigate the first two EOFs.

5.2.2. Metric Comparison

For a reliable spatial model validation, it is inevitable to understand the selected performance metrics. Several spatial performance metrics are compared in Figure 6. The figure depicts the temporal variability of the given metrics for the 33 LST map comparisons and their correlation based on the individual similarity scores for the 33 map comparisons is given in Table 5. All given metrics are equipped with an internal variation, thus there exists sensitivity toward poor and good map agreements. The FK similarity scores are given for the "FK survey" scenario with calibrated λ parameters in the distance decay functions of FoC and FoL and

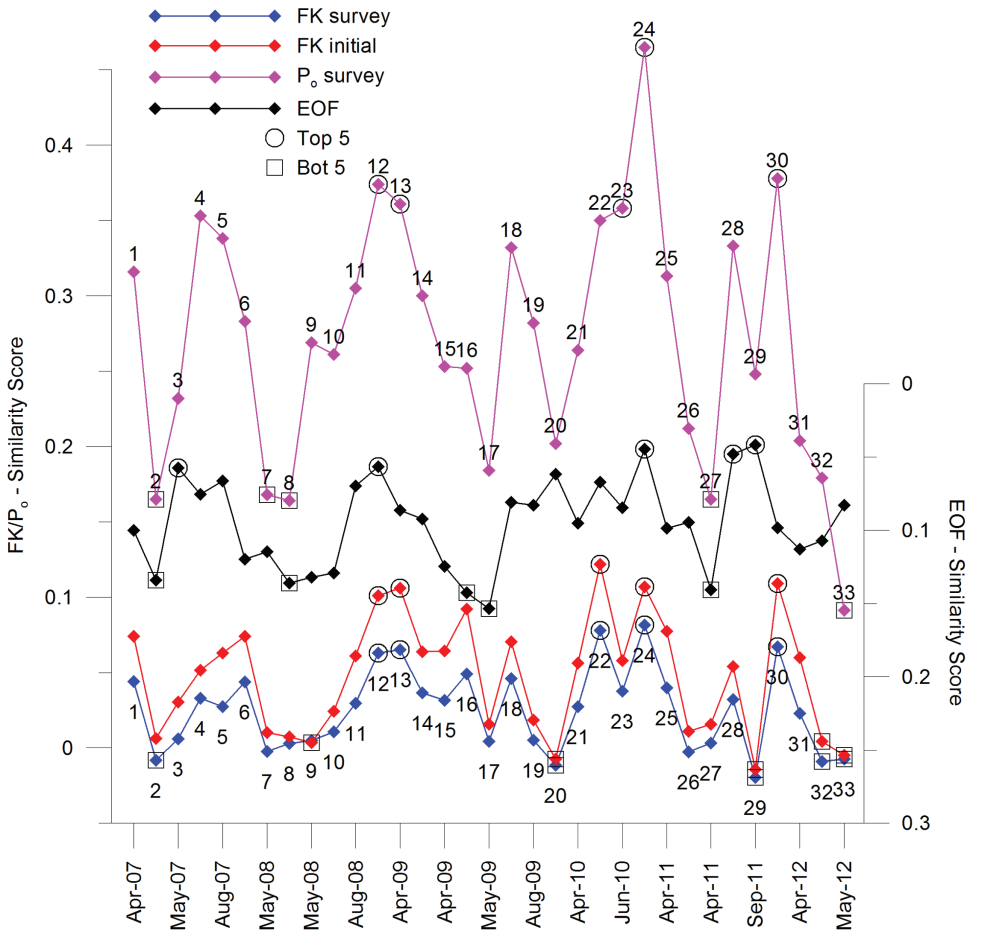

Figure 6. Temporal variation of various metrics: Fuzzy Kappa (λ values after initial guess and λ values calibrated against the survey ranking, equation (5)), observed agreement (λ values calibrated against the survey ranking, equation (5)), and EOF-analysis (reflecting the weighted loading differences of all EOFs, equation (3)). The x axis represents dates of the 33 LST and is thus not equidistant. The top and bottom five map comparisons are underlined by circles and squares, respectively, in order to further analyze the extremes to understand the metric's sensitivity.

the "FK initial" scenario where a qualified guess defines λ. The observed agreement (P_0) originates from the FK equation (equation (4)) and is based on the virtual confusion matrix (Table 2) with FoL and FoC following the "FK survey" scenario. The EOF similarity score considers the weighted loading deviations of all EOFs between the simulated and observed maps (equation (3)) where low values indicate a good agreement. Differences between "FK survey" and "FK initial" are marginal (correlation of 0.98, Table 5), although there is a bias toward higher values for the latter. The accordance between the similarity scores computed by FK and EOF is limited in Figure 6. The conclusion gets further supported by a distinct weak correlation of −0.25 (Table 5) between FK and EOF. Only considering the observed agreement (P_0) seems to be a compromise with an increased correlation to the EOF series of −0.52 and at the same time keeping a strong correlation of 0.83 (Table 5) to the FK series. Figure 7 exemplifies three map comparisons to get a better understanding of the previously presented time series of similarity scores computed (Figure 6). It is noticeable that the model produces less smooth and more scattered LST maps compared to the satellite data. Therefore, the general pattern, represented by the rough allocation of warm and cold areas, and not single cells is of main interest. A simple visual comparison in combination to scatterplots and general cell-by-cell statistics reveal two satisfying (Figures 7b and 7c) and one poor (Figure 7a) LST simulation. All three examples underline that the simulated temperature range is too extreme and that high temperatures are generally simulated too cold. The spatial performance metrics in Figure 6 perform quite differently and are partly contradictory which is emphasized by the three map comparison examples in Figure 7. Comparison of Figures 7a and 7b (2 and 24 in Figure 6, respectively) are very clear (poor and good agreement attested by all metrics,

Table 5. Correlation Matrix of Various Metrics[a]

	R	RMSE	FK	EOF	abs(ME)	P_o	FS
R	1						
RMSE	−0.59	1					
FK	0.47	−0.74	1				
EOF	−0.56	0.66	−0.25	1			
abs(ME)	−0.11	0.71	−0.71	0.03	1		
P_o	0.42	−0.91	0.83	−0.52	−0.76	1	
FS	0.33	−0.82	0.88	−0.29	−0.85	0.95	1

[a]Correlation-coefficient (R), Root-mean-squared-error (RMSE), Fuzzy Kappa, calibrated against the survey ranking (FK), EOF-analysis (EOF), mean-absolute-error (abs(ME)), observed agreement (P_o), equation (6), and the average of the Fuzzy similarity maps (FS).

respectively); comparison Figure 7c (29 in Figure 6) is very ambiguous (EOF high-, FK low-, P_o medium-agreement). This implies that the three metrics have individual sensitivities.

5.2.3. Metrics as Diagnostic Tool

Figure 6 implies that the goodness of fit between observed and simulated LST patterns varies throughout the simulation period. The metrics can assist in learning about the distributed model by serving as a diagnostic tool to identify systematic spatial model deficiencies. The maps of EOF1, EOF2, and mean fuzzy similarity (FS) (Figures 8a–8c) can support the modeler to diagnose systematic spatial model errors. EOF1 and EOF2 decompose the space-time LST data set of simulated and observed maps into similarities and dissimilarities, whereas the FS captures both at the same time. A systematic model deficiency can be anticipated through a visual comparison between the mean error of all 33 LST maps with the average simulated depth to groundwater table (GWT) (Figure 8f and 8i), which reveals a correlation of mean error and depth to GWT. The deficiency might be caused by an over emphasized coupling between surface/subsurface and atmosphere at areas with a shallow GWT and constant water availability, which leads to an overestimation of evapotranspiration and results in a cooling of the land surface. It is desirable that

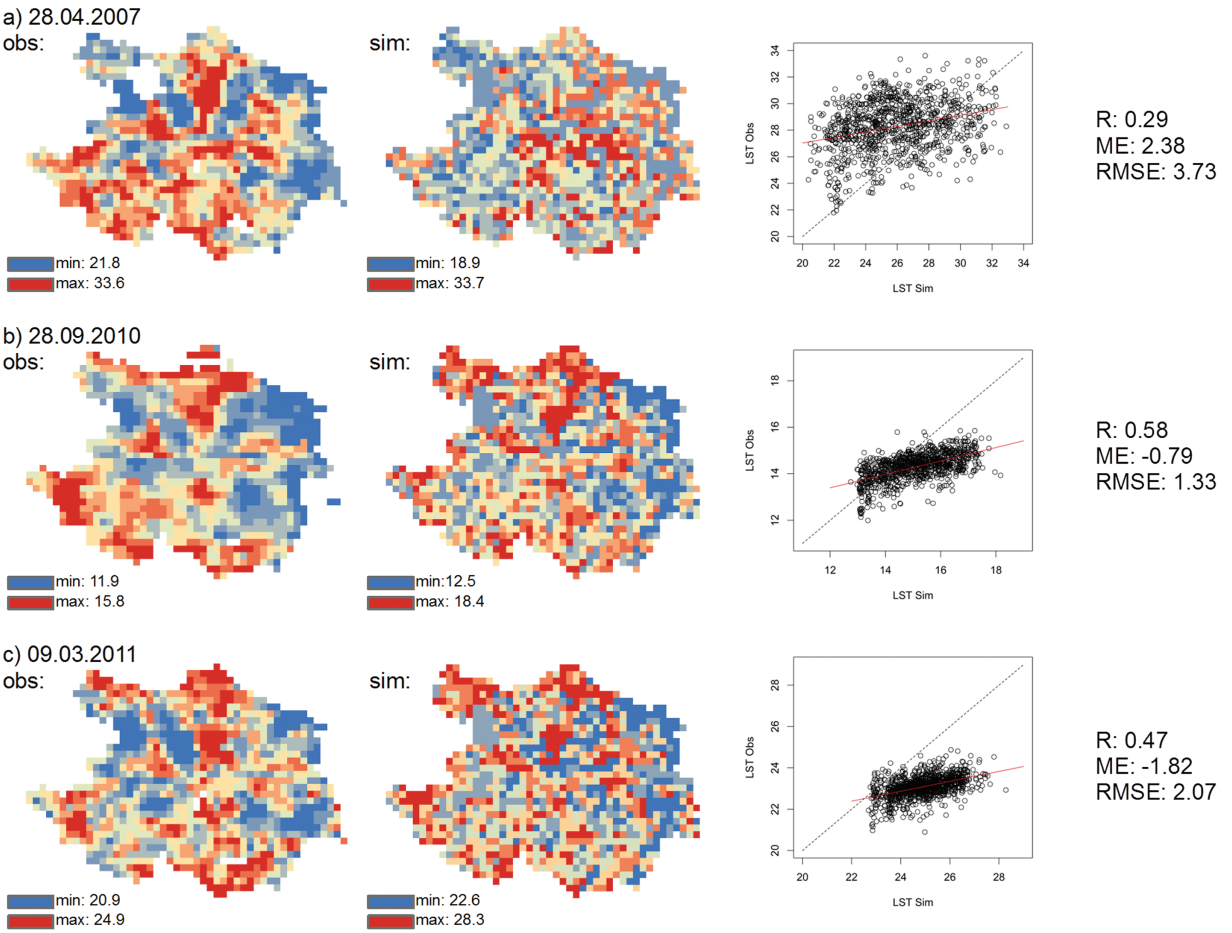

Figure 7. Three examples of simulated and observed temperature maps, with units in degree Celsius. In order to support the visual comparison, scatterplot and basic statistics are given: Correlation-coefficient (R), mean-error (ME), and root mean-squared error (RMSE). The three dates (a), (b), and (c) refer to 2, 24, and 29 in Figure 6.

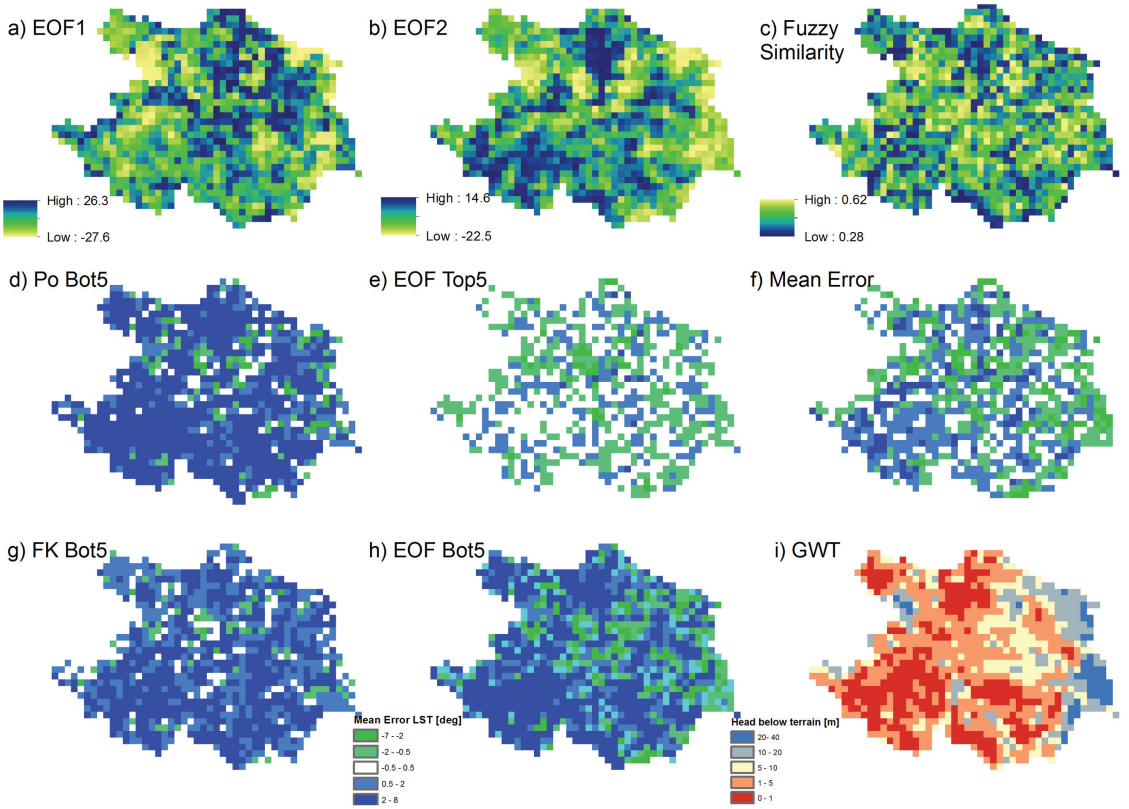

Figure 8. Unitless EOF maps in (a) and (b). (c) Depicts the mean Fuzzy Similarity map (calibrated against survey ranking) based on all 33 LST map comparisons. The mean LST error derived from the bottom five map comparisons indicated by P_o, FK, and EOF-analysis are plotted in (d), (g), and (h), respectively. (e) The mean LST error from the top five map comparisons following the EOF-analysis results. The total mean error is symbolized in Figure 8f and the average simulated depth to groundwater table is presented in Figure 8i.

this deficiency is reflected by the metrics that asses the spatial performance of the model. In order to further investigate the capability of the three above mentioned maps (EOF1, EOF2, and FS) to detect a possible model deficiency, they are plotted in Figure 9a mean-error and Figure 9b depth to GWT. All three maps manifest a "V-shape" when plotted against the mean error with the minimum and maximum around mean error equal to zero for EOF and FS, respectively. This correlation is a good agreement as cell-wise calculated FS attests the highest similarity at cells with a low mean error. The shape of the two EOF plots reflects that both EOF1 and EOF2 are indicators for model errors, where large EOF values are associated with large model errors. This is reasonable because the EOF method is mean corrected, hence it is to be expected that cells with a small temperature anomaly from the mean LST show small errors as well as small EOF values. One has to consider that the sign of an EOF value has no physical relevancy, therefore absolute values are plotted. In the original data, EOF1 and EOF2 have a strong negative and positive correlation to the mean error, respectively. The error bars represent one standard deviation and hence underline the uncertainty of the trendlines due to a large data spread. EOF2 has the lowest uncertainty when plotted against the mean LST error, which strengthens that EOF2 captures the predominant dissimilarities between observed and simulated LST maps. Plotting the data against the average GWT reveals a different picture (Figure 9b). EOF2 still shows a clear "V-shape" correlation, which is supported by a visual comparison of the two maps in Figures 8b and 8i, although the uncertainty increases drastically for deeper GWT values. On the other hand, EOF1 is clearly not correlated to GWT, although a strong correlation to the mean error is prominent, thus it must reflect a different error source. FS does not disaggregate the error sources and thus shows an extreme weak correlation to GWT, the "V-shape" is hardly noticeable and less distinct than in Figure 9a). The trend line representing the relationship between EOF2 and depth to GWT has its minimum (EOF=0) around 5 m of depth to groundwater table. Taking the large uncertainties into account, one can conclude that areas with a moderate depth of 5 m to the groundwater table and above are least affected by the systematic model deficiency. Grids with a shallower depth to GWT are with a higher certainty more affected by the

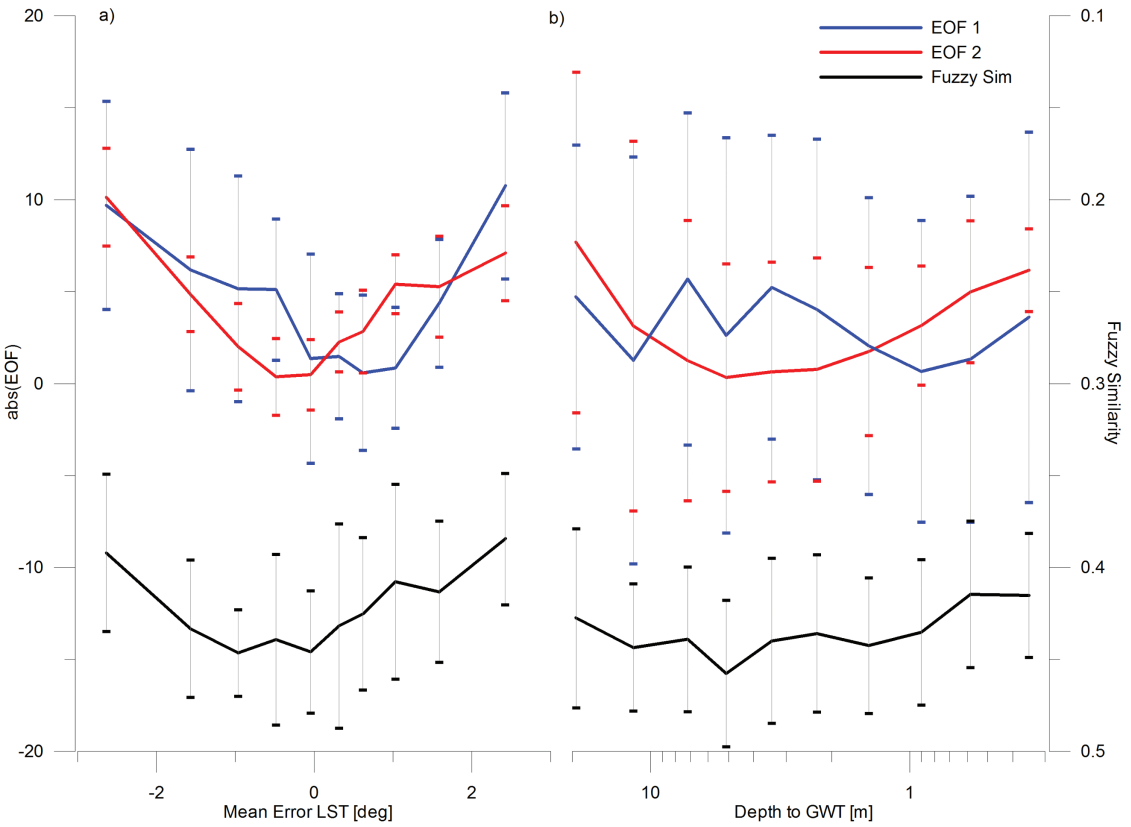

Figure 9. The maps of EOF1, EOF2, and FS (based on survey calibration—high values reflect high similarity) map are plotted against (a) the mean LST error (obs-sim) and (b) the averaged simulated depth to groundwater table (GWT) in meter below terrain. The trend lines are plotted following the description in Figure 5 and are additionally equipped with error bars that represent 1 standard deviation.

systematic underestimation of LST (Figure 9). Opposed and despite an increased uncertainty, deeper GWT cells show a trend toward overestimation of LST. The latter is likely due to a compensating effect in the prior calibration to reach an overall small bias in LST. The EOF-analysis proves the GWT-related model deficiency because the pattern of EOF2 (Figure 8b), which mainly captures the dissimilarities, strongly resembles the average simulated depth to GWT map (Figure 8i). The FS map (Figure 8c) that does not decompose the space-time LST data set into similarities and dissimilarities shows a less distinct but sill noticeable resemblance to the depth to GWT map.

5.2.4. Metric Sensitivities

After successful identification of the major model deficiency via spatial model evaluation, it is of interest to subsequently investigate the sensitivity of the presented metrics to the given GWT-related model defect. This will contribute to the understanding of the strengths and weaknesses of the metrics. In order to examine the predominant sensitivities, mean error maps based on the five map comparisons with the highest (Top5) and lowest (Bot5) similarity scores following Figure 6 for the three metrics observed agreement (P_o), FK and EOF are given in Figures 8d, 8e, 8g, and 8h. This analysis is conducted under the assumption that the Bot5 map comparisons are expected to reflect the identified model deficiency; if not, then the metric must have a different sensitivity. Although the three metrics identify partly different map comparisons for the Top5 category (Figure 6), the derived Top5 maps show a distinct resemblance. Therefore, only the Top5 mean error map based on the EOF similarity score is included in Figure 8e. The Bot5 mean error maps are given for all three metrics as they differ more significantly. A first visual comparison between the Bot5 maps (Figures 8d, 8g, and 8h) and the average simulated depth to GWT map (Figure 8i) leads to the conclusion that the EOF similarity score performs best at identifying the main model deficiency. Moreover, the GWT correlation is addressed in Figure 10. Two Top5 maps (FK and P_o) are included in Figure 10 to underline their similarity. Both of them plot close to a mean LST error of zero with small uncertainties. The metrics

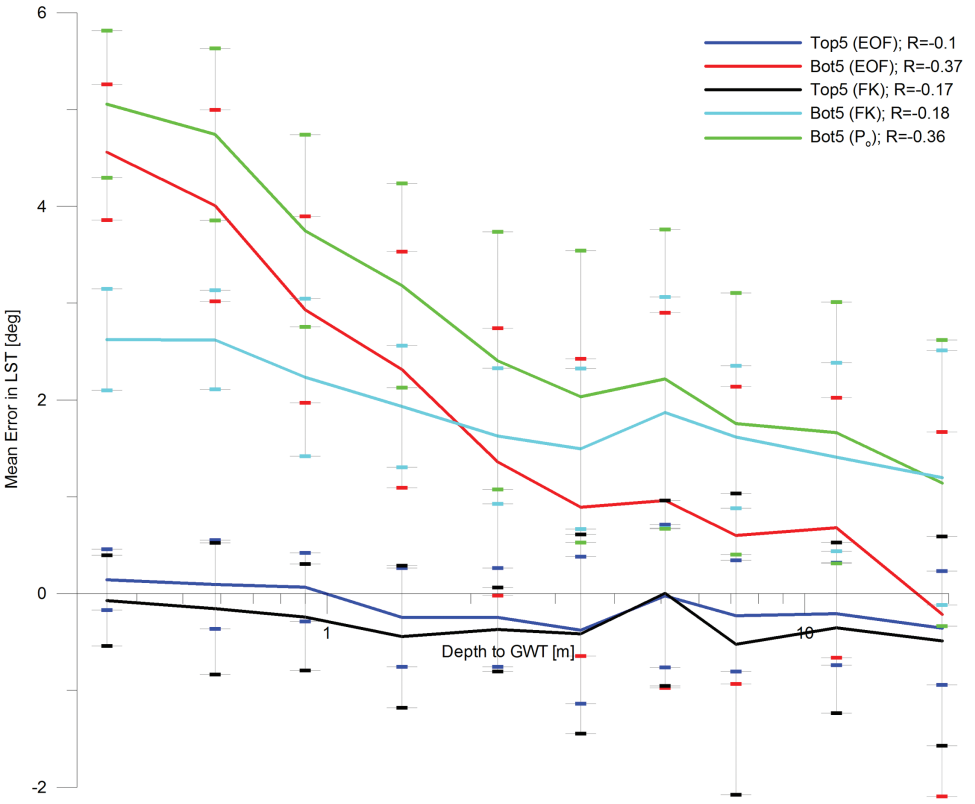

Figure 10. The extremes (mean error maps derived by the bottom- and top-5 maps) of (a) EOF-analysis and (b) Fuzzy Kappa (calibrated against survey ranking) are plotted against the average simulated groundwater table in meter below terrain. R indicates the correlation between mean error to GWT. The trend lines are plotted following the description in Figure 5 and are additionally equipped with error bars that represent 1 standard deviation.

EOF, FK, and P_o perform differently in identifying the LST map comparisons with the lowest similarity score (Bot5) as shown in Figure 10. The correlation between GWT and mean LST error of Bot5 identified by the EOF similarity score is most profound as the mean error increases drastically at cells with a shallow depth to GWT (comparing Figures 8h and 8i). FK misses the GWT correlation, instead the Bot5 map reflects a general positive bias. The pronounced bias sensitivity of FK is found to be less apparent for the observed agreement (P_o) which is a part of the FK equation (equation (4)). The GWT correlation for Bot5-P_o is less distinct than for Bot5-EOF as the mean error (Bot5-P_o) is still high at cells with a deep GWT (Figure 10). The bias sensitivity of FK lies in its equation because the observed agreement (P_o) is corrected by the expected agreement (P_e). If the latter is larger than the first, FK yields negative values (equation (4)). Figure 11 underlines this and also shows the dependency of FK to the model bias. With increasing model bias, increases the chance that the expected agreement (P_e) grows larger than the actual observed agreement P_o resulting in a negative FK. This effect is apparent where the defined FoC (Table 2) loses influence; around $\pm 2^\circ$. Both, P_e and P_o are established on the same virtual confusion matrix. Considering equations (6) and (7) stresses that P_e is based on the row and column sums only, thus not depending on the actual positioning of the cells. Opposed, P_o uses the single entries of the virtual confusion matrix to compute the agreement. This difference is fundamental because it makes P_e more resilient to a bias than P_o.

6. Discussion

6.1. Survey

There is no common understanding about the sensitivities and peculiarities of spatial performance metrics. Therefore, the human eye survey proved as a very reliable source to test and benchmark the metrics as well as to calibrate the λ values in the distance decay function (equation (5)) for FK. The survey design as well as the way the synthetic maps are perturbed remain of course highly subjective and are thus disputable. More

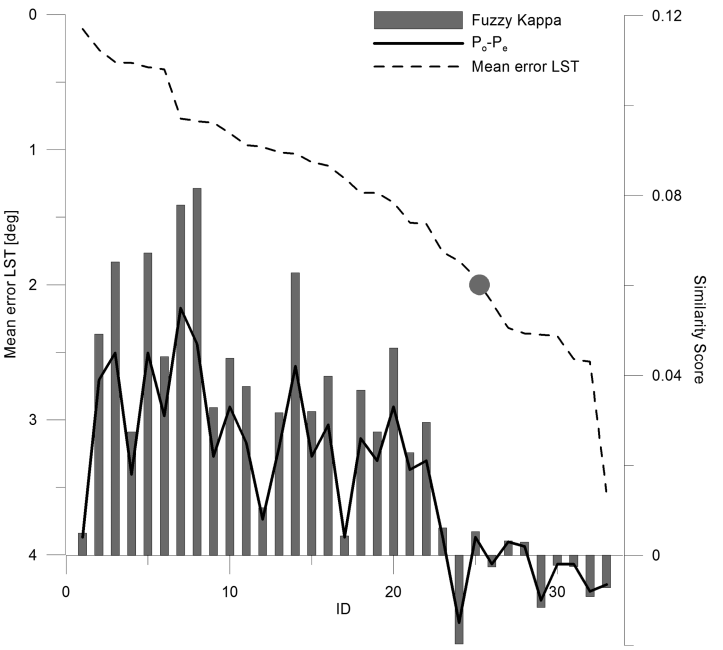

Figure 11. The bias sensitivity of FK is underlined by plotting the similarity score derived by FK in dependency on mean LST per map comparison. The 33 map comparisons are sorted after an increasing mean LST error. The observed agreement (Po)—expected agreement (Pe), both from equation (4), highly affects the computed similarity score. Mean LST error of 2 degrees marks the point where FoC becomes insignificant (Table 4).

generally, there are two main points to debate. First, it is not assured if the human perception is a credible source to quantify map comparisons in the context of hydrological modeling. However, without being experts in hydrology, it can be argued that humans are well trained in pattern recognition and comparison. Second, a metric that performs satisfyingly in reproducing the survey ranking is not automatically suitable for a model evaluation or calibration. This is mainly because the survey probably does not represent the full magnitude and interconnection of possible sources of model errors. The survey clearly assists in identifying sensitivities and peculiarities of the spatial metrics. The EOF-analysis ranks ID#4 and ID#11 (perturbed by spatial shift) with a bad similarity score because the metric underlies not a true geostatistical approach and hence it is very sensitive for uncertainties related to location (Table 3 and Figure 3). Along those lines, the FK method facilitates freedom to focus on uncertainty in location by pronouncing FoL, which yields a relatively higher similarity score of the two given maps in the FK LOC scenario (Table 3). The bias perturbed map ID#10 is most peculiar because the EOF-analysis is completely bias insensitive, whereas FK manifest a strong bias sensitivity, unless FoC is overly pronounced. The survey ranking is utilized to calibrate parameters in the distance decay functions to derive FoL and FoC. However, the initially chosen λ values in the distance decay function (equation (5)) perform surprisingly well in reproducing the ranking compared to the calibrated case (Table 3) as well as in assigning the similarity scores in Figure 6.

6.2. Kappa and Fuzziness

Any application of Kappa statistics or fuzzy map comparison requires a categorization of the variable of interest. Therefore, it is very suitable for spatial evaluation of more discrete variables like e.g., snow-cover or precipitation. One degree intervals for the categorization of the LST maps appear adequate to represent general temperature patterns because uncertainty of the observation may exceed 1° and the correct simulation of very precise temperature values is not the focus of this study. This procedure could easily be applied to other continuous variables, e.g., precipitation. The definition of the distance decay function for the assessment of FoL and FoC is highly subjective; especially the threshold which marks the boundary (maximum distance in meter and category, respectively) of membership is difficult to interpret. The threshold for FoL can be assessed by a variogram analysis of the data with focus on the correlation length. The definition of the FoC threshold is more peculiar. However, knowledge on uncertainties of the observed data or on the accepted uncertainty of the model output can give guidance. This study promotes the Fuzzy Weighted Kappa [*Huang and Lees*, 2007] as FK because of its easy application. Moreover, it is preferred over other Fuzzy Kappa approaches, like e.g., *Hagen et al*. [2005] because it is more applicable for an accuracy assessment, comparing neighborhood to cell instead of neighborhood to neighborhood, which is implied by the two-way comparison in the Fuzzy Kappa definition by *Hagen et al*. [2005]. More precisely, this study only considers fuzziness in the simulated LST data and not in the observed satellite data. Taking fuzziness of both data sets simultaneously into account is in theory possible, but would mix the model evaluation with

an uncertainty assessment of the observation data. The latter is extremely important, but should be dealt with beforehand to keep the validation transparent. A recent discussion driven by *Pontius and Millones* [2011] "Death to Kappa" and *van Vliet et al*. [2013] "Rebirth of Kappa" underlines the ongoing dispute on Kappa statistics. The discussion is mainly in the context of remote sensing and land use modeling. However, the main controversy, namely using the expected agreement (P_e) following randomness as a baseline in the Kappa equation (equation (4)), proves to be very relevant in this hydrological modeling study. Standardizing by P_e strengthens the bias sensitivity of Kappa, which is stressed by the low ranking of the bias perturbed map ID#10 in the survey (Table 3), by Figure 11 and in addition by the strong correlation between bias and FK in Table 5. Both above mentioned studies promote innovative baselines for P_e; however, none emerges as practical and easy to apply in hydrology because they are drawn from applications in the fields of land use change modeling and remote sensing. On the contrary of *van Vliet et al*. [2013] concerns about not using a baseline at all, the solitary observed agreement P_o proves very adequate in this study (Figure 10 and Table 5). P_o gives a compromise between the pronounced bias sensitivity of FK and the bias insensitivity of the EOF-analysis.

6.3. EOF-Analysis

Applying the EOF-analysis as a spatial evaluation tool by comprising both, the model output and the observational data into the time-space matrix is a novel approach. This study clearly promotes this approach as a diagnostic tool to detect model deficiencies in hydrological modeling because the EOF-analysis decomposes the spatial model evolution into maps of similarities and dissimilarities. The EOF-analysis as well as simple statistics like RMSE, mean error, or correlation coefficient are essentially a cell-by-cell comparison and can thus not be defined as a true spatial performance metric as no information on spatial correlation of the patterns is incorporated. Nevertheless, the EOF-analysis can be extended by a variogram analysis of the EOFs to overcome the geostatistical limitations [*Graf et al*., 2012]. The final EOF-based similarity score comprises the weighted loading differences between simulated and observed maps of all EOFs. The associated weights to the loading deviations in equation (3) are necessary, but are however also misleading because they represent both the intervariability and intravariability of the observed and simulated LST data set. The interest of the spatial performance metric is solely focused on the intervariability, so new ways to weigh the loading deviations must emerge. Alternatively, if only the main error source is of interest and minor errors are negligible, only the loadings deviations of the second EOF could be assessed. Generally the proposed EOF-method could be applied to other spatial variables in the context of hydrological modeling e.g., the comparison of simulated and observed soil moisture patterns in large-scale hydrological modeling [*Pan et al*., 2014] or the quantitative comparison of various precipitations forecasts [*Surcel et al*., 2014]. The latter might be more discrete in time and space compared to the LST data which has smooth patterns in time and space. However, discreetness will not restrict the applicability of the EOF analysis because the sequence of rows and columns in the time-space input matrix is indifferent. The applicability of the EOF-analysis on snow-cover data [*Duethmann et al*., 2014] is expected to be limited because the EOF method is unsuited for binary data.

6.4. Metrics as a Diagnostic Tool

Similarity may have different interpretations and a reflection on this is necessary before choosing a spatial performance metric. In many model applications a distinct bias, while retaining a correct simulation of the spatio-temporal dynamics is acceptable. The bias can then either be reduced through inverse calibration or just be accepted due to differences in scales between observation and model. The perturbation with ID#10 in the survey represents a clear bias while keeping the overall pattern of the reference map. Following Table 3, only R and EOF rate ID#10 with the highest similarity, while FK is bias sensitive and attest an overall poor similarity to ID#10. Due to different sensitivities all presented metrics correlate differently (Table 5), which has to be noted and can be utilized by the modeler in selecting the right performance metric. Supposedly one single performance metric is not sufficient for a thorough spatial model evaluation, which was already argued by *Wealands et al*. [2005]. Therefore, a combination of two or more metrics it possible: E.g., the EOF-analysis to assess the general patterns in combination with the mean error to account for the model bias.

The FS- and the EOF-maps add a spatial dimension to the spatial model evaluation and prove very instructive as diagnostic tools. Both pinpoint the same model deficiency, namely the over-emphasized coupling in the SW-ET model which yields undersimulated LST due to cooling caused by too high evapotranspiration in

grids cells with a shallow depth to GWT. The FS map shows a vague relation to this model deficiency because it also contains other sources of error, such as bias. The EOF-analysis, in contrast, unambiguously identifies the deficiency because it decomposes the similarities and various sources of errors between the simulated and observed LST maps. *Yoo and Kim* [2004] make use of the EOF decomposition in a different context, where they successfully correlate EOFs with landscapes characteristics, such as elevation slope or porosity to identify the main contributor to spatial variability in soil moisture at catchment scale. Analyzing and solving the identified model deficiency must be scope of a follow-up study. Apart from interpreting the spatial distribution of the LST error it should be aimed at an overall similarity score for an objective and robust spatial model evaluation, which could be applied in an inverse model calibration. This is clearly provided by FK and P_o and can be derived for the EOF analysis by summing the weighted loading differences. The average of each individual FS map, giving an overall accuracy measure, is highly correlated to P_0 (Table 5) therefore the application of both seems redundant.

7. Conclusions

There is a pertinent need to measure the predictive accuracy of spatial outputs of distributed hydrological models and development and testing of new metrics for this purpose are required. Until now there are no standard performance metrics for evaluating simulated spatial patterns against an observed pattern such as there are standards to assess the models efficiency in runoff simulation by means of the Nash-Sutcliffe efficiency [*Gupta et al*., 2009]. Therefore, rigorous testing is needed to create standards and a common understanding of spatial model evaluation in hydrology. Spatial model evaluation can be expected to become an important part of multi-objective calibration of distributed hydrological models [*Efstratiadis and Koutsoyiannis*, 2010]. The spatial performance metrics featured by this study show some redundancy. Nevertheless, they facilitate diverse information about the agreement of simulated and observed patterns and are equipped with different sensitivities. One important outcome of this study is that the bias sensitivity turned out to be clearly distinguishable in the tested metrics. The bias sensitivity is most distinct in the Fuzzy Kappa method and unapparent for the EOF-analysis, due to a prior mean correction. The focus of a spatial model evaluation can be to obtain an accurate simulation of patterns in absolute or relative terms or in a balanced manner. The latter is always an attractive option and is well represented by P_o in this study because the bias sensitivity can clearly be regulated by the way FoC is defined. The survey ranking offers the opportunity to benchmark various spatial performance metrics and it should be exploited further by extending Table 3 with further metrics. Within atmospheric sciences *Hering and Genton* [2011] and *Gilleland* [2013] have successfully applied the Spatial Prediction Comparison Test (SPCT) on simulated wind fields and precipitation forecasts, respectively. SPCT considers the spatial correlation structure through empirical semivariogram and does not impose a distributional assumption of the continuous variable of interest. Another promising method builds up on advances in wavelet analysis in hydrology [*Labat*, 2005] that allow new insight into the scale dependency of a signal which can be either temporal (1D) or spatial (2D). In more detail a continuous wavelet coherence analysis investigates the scale- and location- dependency of two signals, which is utilized by *Si and Zeleke* [2005] and [*Si* [2008]] to conduct a scaling analysis of various soil physical properties to saturated hydraulic conductivity. Both approaches can be applied to assess the similarities between two LST maps.

Acknowledgments:
The work has been carried out under the HOBE (Center for Hydrology) and SPACE (SPAtial Calibration and Evaluation in distributed hydrological modeling) project, both founded by the VILLUM Foundation. The used land surface temperature maps, for the survey and for the modeling case study will be made publicly available on the first author's research gate profile as supporting materials to this publication. Alternative the first author can be contacted directly via mail (juko@geus.dk) for a data request. Further we would like to acknowledge Alexander Graf from the Agrosphere Institute, Forschungszentrum Jülich for his help on the development and the analysis of the EOF methodology for pattern comparison.

References

Abbott, M. B., J. C. Bathurst, J. A. Cunge, P. E. Oconnell, and J. Rasmussen (1986), An introduction to the European hydrological system: Systeme hydrologique Europeen, she .2. Structure of a physically-based, distributed modeling system, *J. Hydrol.*, *87*(1–2), 61–77.

Ahmad, I., R. Ambreen, S. Sultan, Z. Sun, and W. Deng (2014), Spatial-temporal variations in January precipitation over the period of 1950–2000 in Pakistan and possible links with global teleconnections: Geographical perspective, *Am. J. Clim. Change*, *3*(4), 378–387.

Bennett, N. D., et al. (2013), Characterising performance of environmental models, *Environ. Modell. Software*, *40*, 1–20.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*(3), 279–298.

Beven, K., and J. Feyen (2002), The future of distributed modelling: Special issue, *Hydrol. Processes*, *16*(2), 169–172.

Brown, B. G., E. Gilleland, and E. E. Ebert (2011), Forecasts of spatial fields, in *Forecast Verification*, edited by I. T. Jolliffe, and D. B. Stephenson, pp. 95–117, John Wiley, Hoboken, N. J.

Brown, B. G., J. H. Gotway, R. Bullock, E. Gilleland, T. Fowler, D. Ahijevych, and T. Jensen (2009), The Model Evaluation Tools (MET): Community tools for forecast evaluation, paper presented at Preprints, 25th Conference on International Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Phoenix, AZ, Amer. Meteor. Soc. A., American Meteorological Society, Boston, Mass.

Carterette, E., and M. Friedman (1974), *Handbook of Perception: Historical and Philosophical Roots of Perception*, vol. 1, 41–55, Academic, N. Y.
Cohen, J. (1960), A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.*, *20*(1), 37–46.
Cohen, J. (1968), Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit, *Psychol. Bull.*, *70*(4), 213–220.
Cornelissen, T., B. Diekkrüger, and H. R. Bogena (2014), Significance of scale and lower boundary condition in the 3D simulation of hydrological processes and soil moisture variability in a forested headwater catchment, *J. Hydrol.*, *516*(0), 140–153.
Duethmann, D., J. Peters, T. Blume, S. Vorogushyn, and A. Güntner (2014), The value of satellite-derived snow cover images for calibrating a hydrological model in snow-dominated catchments in Central Asia, *Water Resour. Res.*, *50*, 2002–2021, doi:10.1002/2013WR014382.
Efstratiadis, A., and D. Koutsoyiannis (2010), One decade of multi-objective calibration approaches in hydrological modelling: A review, *Hydrol. Sci. J.-Journal Des Sciences Hydrologiques*, *55*(1), 58–78.
Foody, G. M. (2008), Harshness in image classification accuracy assessment, *Int. J. Remote Sens.*, *29*(11), 3137–3158.
Fritz, S., and L. Lee (2005), Comparison of land cover maps using fuzzy agreement, *Int. J. Geogr. Inf. Sci.*, *19*(7), 787–807.
Gilleland, E. (2013), Testing competing precipitation forecasts accurately and efficiently: The spatial prediction comparison test, *Mon. Weather Rev.*, *141*(1), 340–355.
Gilleland, E., D. A. Ahijevych, B. G. Brown, and E. E. Ebert (2010), Verifying forecasts spatially, *Bull. Am. Meteorol. Soc.*, *91*(10), 1365–1373.
Graf, A., M. Herbst, L. Weihermüller, J. A. Huisman, N. Prolingheuer, L. Bornemann, and H. Vereecken (2012), Analyzing spatiotemporal variability of heterotrophic soil respiration at the field scale using orthogonal functions, *Geoderma*, *181–182*(0), 91–101.
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. Bloschl, A. W. Western, and T. A. McMahon (2002), Advances in the use of observed spatial patterns of catchment hydrological response, *Adv. Water Resour.*, *25*(8–12), 1313–1334.
Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*(1–2), 80–91.
Guzinski, R., H. Nieto, S. Stisen, and R. Fensholt (2014), Inter-comparison of energy balance and hydrological models for land surface energy fluxes estimation over a whole river catchment, *Hydrol. Earth Syst. Sci. Discuss.*, *11*(6), 5905–5951.
Hagen, A. (2003), Fuzzy set approach to assessing similarity of categorical maps, *Int. J. Geogr. Inform. Sci.*, *17*(3), 235–249.
Hagen, A. (2009), An improved Fuzzy Kappa statistic that accounts for spatial autocorrelation, *Int. J. Geogr. Inform. Sci.*, *23*(1), 61–73.
Hagen, A., and P. Martens (2008), Map comparison methods for comprehensive assessment of geosimulation models, paper presented at ICCSA (1), 194–209, Springer Berlin Heidelberg.
Hagen, A., B. Straatman, and I. Uljee (2005), Further developments of a fuzzy set map comparison approach, *Int. J. Geogr. Inform. Sci.*, *19*(7), 769–785.
Hannachi, A., I. T. Jolliffe, and D. B. Stephenson (2007), Empirical orthogonal functions and related techniques in atmospheric science: A review, *Int. J. Climatol.*, *27*(9), 1119–1152.
Hering, A. S., and M. G. Genton (2011), Comparing spatial predictions, *Technometrics*, *53*(4), 414–425.
Huang, Z., and B. G. Lees (2007), Assessing a single classification accuracy measure to deal with the imprecision of location and class: Fuzzy weighted Kappa versus Kappa, *J. Spatial Sci.*, *52*(1), 1–12.
Jacob, P., and M. Jeannerod (2003), *Ways of Seeing: The Scope and Limits of Visual Cognition (Oxford Cognitive Science)*, Oxford Univ. Press.
Jawson, S. D., and J. D. Niemann (2007), Spatial patterns from EOF analysis of soil moisture at a large scale and their dependence on soil, land-use, and topographic properties, *Adv. Water Resour.*, *30*(3), 366–381.
Jensen, K. H., and T. H. Illangasekare (2011), HOBE: A hydrological observatory, *Vadose Zone J.*, *10*(1), 1–7.
Korres, W., C. N. Koyama, P. Fiener, and K. Schneider (2010), Analysis of surface soil moisture patterns in agricultural landscapes using Empirical orthogonal functions, *Hydrol. Earth Syst. Sci.*, *14*(5), 751–764.
Kuhnert, M., A. Voinov, and R. Seppelt (2005), Comparing raster map comparison algorithms for spatial modeling and analysis, *Photogramm. Eng. Remote Sens.*, *71*(8), 975–984.
Labat, D. (2005), Recent advances in wavelet analyses: Part I. A review of concepts, *J. Hydrol.*, *314*(1–4), 275–288.
Liu, T., P. Willems, X. W. Feng, Q. Li, Y. Huang, A. M. Bao, X. Chen, F. Veroustraete, and Q. H. Dong (2012), On the usefulness of remote sensing input data for spatially distributed hydrological modelling: Case of the Tarim River basin in China, *Hydrol. Processes*, *26*(3), 335–344.
Müller, B., M. Bernhardt, and K. Schulz (2014), Identification of catchment functional units by time series of thermal remote sensing images, *Hydrol. Earth Syst. Sci. Discuss.*, *11*(6), 7019–7052.
Overgaard, J. (2005), Energy-based land-surface modelling: New opportunities in integrated hydrological modelling, PhD thesis, Tech. Univ. of Denmark, Inst. of Environ. and Resour, Kongens Lyngby, Denmark.
Pan, M., A. K. Sahoo, and E. F. Wood (2014), Improving soil moisture retrievals from a physically-based radiative transfer model, *Remote Sens. Environ.*, *140*(0), 130–140.
Perry, M. A., and J. D. Niemann (2007), Analysis and estimation of soil moisture at the catchment scale using EOFs, *J. Hydrol.*, *334*(3–4), 388–404.
Pokhrel, P., and H. V. Gupta (2011), On the ability to infer spatial catchment variability using stream flow hydrographs, *Water Resour. Res.*, *47*, W08534, doi:10.1029/2010WR009873.
Pokhrel, P., K. K. Yilmaz, and H. V. Gupta (2012), Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures, *J. Hydrol.*, *418*, 49–60.
Pontius, R. G., and M. Millones (2011), Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment, *Int. J. Remote Sens.*, *32*(15), 4407–4429.
Pontius, R. G., D. Huffaker, and K. Denman (2004), Useful techniques of validation for spatially explicit land-change models, *Ecol. Modell.*, *179*(4), 445–461.
Power, C., A. Simms, and R. White (2001), Hierarchical fuzzy pattern matching for the regional comparison of land use maps, *Int. J. Geogr. Inform. Sci.*, *15*(1), 77–100.
Refsgaard, J. C. (1997), Parameterisation, calibration and validation of distributed hydrological models, *J. Hydrol.*, *198*(1–4), 69–97.
Refsgaard, J. C. (2000), Towards a formal approach to calibration and validation of models using spatial data, in *Spatial Patterns in Catchment Hydrology*, edited by R. Grayson and G. Bloeschl, pp. 329–354, Cambridge Univ. q.
Refsgaard, J. C., et al. (2014), Nitrate reduction in geologically heterogeneous catchments: A framework for assessing the scale of predictive capability of hydrological models, *Sci. Total Environ.*, *468*, 1278–1288.

Remmel, T. K. (2009), Investigating global and local categorical map configuration comparisons based on coincidence matrices, *Geogr. Anal.*, *41*(2), 144–157.
Rose, K. A., B. M. Roth, and E. P. Smith (2009), Skill assessment of spatial maps for oceanographic modeling, *J. Mar. Syst.*, *76*(1-2), 34–48.
Sciuto, G., and B. Diekkruger (2010), Influence of soil heterogeneity and spatial discretization on catchment water balance modeling, *Vadose Zone J.*, *9*(4), 955–969.
Shuttleworth, W. J., and J. S. Wallace (1985), Evaporation from sparse crops: An energy combination theory, *Q. J. R. Meteorol. Soc.*, *111*(469), 839–855.
Si, B. C. (2008), Spatial scaling analyses of soil physical properties: A review of spectral and wavelet methods, *Vadose Zone J.*, *7*(2), 547–562.
Si, B. C., and T. B. Zeleke (2005), Wavelet coherency analysis to relate saturated hydraulic properties to soil physical properties, *Water Resour. Res.*, *41*, W11424, doi:10.1029/2005WR004118.
Silvestro, F., S. Gabellani, F. Delogu, R. Rudari, and G. Boni (2013), Exploiting remote sensing land surface temperature in distributed hydrological modelling: The example of the Continuum model, *Hydrol. Earth Syst. Sci.*, *17*(1), 39–62.
Smith, M. B., and H. V. Gupta (2012), The distributed model intercomparison project (DMIP): Phase 2 experiments in the Oklahoma Region, USA, *J. Hydrol.*, *418–419*(0), 1–2.
Smith, M. B., V. Koren, Z. Zhang, Y. Zhang, S. M. Reed, Z. Cui, F. Moreda, B. A. Cosgrove, N. Mizukami, and E. A. Anderson (2012), Results of the DMIP 2 Oklahoma experiments, *J. Hydrol.*, *418–419*(0), 17–48.
Spillmann, L., and J. S. Werner (1990), *Visual Perception: The Neurophysiological Foundations*, Academic.
Stisen, S., M. F. McCabe, J. C. Refsgaard, S. Lerer, and M. B. Butts (2011a), Model parameter analysis using remotely sensed pattern information in a multi-constraint framework, *J. Hydrol.*, *409*(1-2), 337–349.
Stisen, S., T. O. Sonnenborg, A. L. Hojberg, L. Troldborg, and J. C. Refsgaard (2011b), Evaluation of climate input biases and water balance issues using a coupled surface-subsurface model, *Vadose Zone J.*, *10*(1), 37–53.
Sun, Z., N.-B. Chang, Q. Huang, and C. Opp (2012), Precipitation patterns and associated hydrological extremes in the Yangtze River basin, China, using TRMM/PR data and EOF analysis, *Hydrol. Sci. J.*, *57*(7), 1315–1324.
Surcel, M., I. Zawadzki, and M. Yau (2014), On the filtering properties of ensemble averaging for storm-scale precipitation forecasts, *Mon. Weather Rev.*, *142*(3), 1093–1105.
van Vliet, J., A. Hagen-Zanker, J. Hurkens, and H. van Delden (2013), A fuzzy set approach to assess the predictive accuracy of land use simulations, *Ecol. Modell.*, *261*, 32–42.
Wagemans, J., J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt (2012), A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization, *Psychol. Bull.*, *138*(6), 1172–1217.
Wang, S. Y., R. McGrath, T. Semmler, and C. Sweeney (2006), Validation of simulated precipitation patterns over Ireland for the period 1961–2000, *Int. J. Climatol.*, *26*(2), 251–266.
Warscher, M., U. Strasser, G. Kraller, T. Marke, H. Franz, and H. Kunstmann (2013), Performance of complex snow cover descriptions in a distributed hydrological model system: A case study for the high Alpine terrain of the Berchtesgaden Alps, *Water Resour. Res.*, *49*(5), 2619–2637.
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*(1), 15–32.
Wertheimer, M. (1912), Experimentelle Studien über das Sehen von Bewegung (Translated extract reprinted as "Experimental studies on the seeing of motion"), *Zeitschrift für Psychologie*, *61*, 161–265.
Yoo, C., and S. Kim (2004), EOF analysis of surface soil moisture field variability, *Adv. Water Resour.*, *27*(8), 831–842.
Zadeh, L. A. (1965), Fuzzy sets, *Inform. Control*, *8*(3), 338.
Zadeh, L. A. (1968), Fuzzy algorithms, *Inform. Control*, *12*(2), 94 pp.