Spatial validation of large scale land surface models against monthly land surface temperature patterns using innovative performance metrics.

Autors:

Julian Koch[1, 2, 3,*], Amanda Siemann[3], Simon Stisen[2] and Justin Sheffield[3]

Affiliations:

1) Department of Geosciences and Natural Resources Management, University of Copenhagen, DK

2) Department of Hydrology, Geological Survey of Denmark and Greenland, DK

3) Department of Environmental and Civil Engineering, Princeton University, USA

Submitted to: Journal of Geophysical Research – Atmospheres, 11th of November 2015

Re-Submitted to: Journal of Geophysical Research – Atmospheres, 29th of March 2016

*Corresponding author address:

Julian Koch - Geological Survey of Denmark and Greenland (GEUS), Department of Hydrology - Øster Voldgade 10, Copenhagen, Denmark - E-mail: juko@geus.dk , Telephone: +45 38142768

Key Points

- Comprehensive spatial validation of three land surface models over the contiguous USA.
- Incorporating a 30 year remote sensing dataset of monthly land surface temperature maps.
- Application of two innovative performance metrics to assess the simulated spatial patterns.

1. Abstract

Land surface models (LSMs) are a key tool to enhance process understanding and to provide predictions of the terrestrial hydrosphere and its atmospheric coupling. Distributed LSMs predict hydrological states and fluxes, such as land surface temperature (LST) or actual evapotranspiration (aET), at each grid cell. LST observations are widely available through satellite remote sensing platforms that enable comprehensive spatial validations of LSMs. In spite of the great availability of LST data, most validation studies rely on simple cell to cell comparisons and thus do not regard true spatial pattern information. The core novelty of this study is the development and application of two innovative spatial performance metrics, namely EOF- and connectivity-analysis, to validate predicted LST patterns by three LSMs (Mosaic, Noah, VIC) over the contiguous USA. The LST validation dataset is derived from global High-Resolution-Infrared-Radiometric-Sounder (HIRS) retrievals for a 30 year period. The metrics are bias insensitive, which is an important feature in order to truly validate spatial patterns. The EOF analysis evaluates the spatial variability and pattern seasonality, and attests better performance to VIC in the warm months and to Mosaic and Noah in the cold months. Further, more than 75% of the LST variability can be captured by a single pattern that is strongly correlated to air temperature. The connectivity analysis assesses the homogeneity and smoothness of patterns. The LSMs are most reliable at predicting cold LST patterns in the warm months and vice versa. Lastly, the coupling between aET and LST is investigated at flux tower sites and compared against LSMs to explain the identified LST shortcomings.

2. Introduction

The terrestrial hydrological cycle comprises a complex interplay of subsurface, surface and atmosphere processes with direct implications for the energy and carbon cycles. Reliably observing and modelling of hydrologic variability and land-atmosphere interactions are a grand scientific and societal challenge addressing issues of e.g. water resources management, climate change, drought and flood risk or land use management. In this regard, distributed land surface modeling is an active field of research that aims at predicting hydrologic variability at catchment scale (e.g. Stisen et al., 2011), large basin scale (e.g. Getirana et al., 2014; Long et al., 2014), continental scale (e.g. Sheffield et al., 2014; Troy et al., 2011) or global scale (e.g. Koirala et al., 2014; Sheffield and Wood, 2007). Due to the distinct spatial heterogeneity of the natural system, the distributed nature of a land surface model (LSM) is essential. This allows a process-based LSM to estimate hydrological states and fluxes as well as energy fluxes at each grid (Clark et al., 2015).

In the hydrological community, models are typically validated against discharge at the outlet of a catchment (Refsgaard, 1997). This traditional validation framework is found to have limited sensitivity to the spatial patterns of spatially explicit hydrological variables, like soil moisture or land surface temperature (LST) (Koch et al., 2015a; Stisen et al., 2011). The utility of LSM predictions for understanding, for example, drought and flood risk, land use change effects, or land-atmosphere feedbacks, is therefore hampered by the uncertainty in the representation of the spatial variability of hydrological states and their related fluxes within a catchment or region. Refsgaard (2001) and Grayson et al. (2002) stressed the need to move away from the traditional paradigm of validating distributed

LSMs against aggregated observations such as discharge to a more adequate framework that includes spatial observational data instead. Satellite remote sensing data provides independent spatial observations of hydrological variables that are often at a similar spatial scale as the model's predictions (Wood et al., 2011) and can thus be used for calibrating LSMs (Corbari and Mancini, 2014; Wanders et al., 2014) or be incorporated in data assimilation studies (Moradkhani, 2008; Reichle et al., 2010).

LST is considered a key state variable that controls energy and water exchanges at the land surface-atmosphere interface (Karnieli et al., 2010; Sun and Pinker, 2003). Spatially continuous LST retrievals are widely available through various remote sensing platforms as presented by Li et al. (2013) and Wan et al. (2002). Gunshor et al. (2004) lists and compares various satellite instruments that measure thermal infrared signatures from the earth's surface, which is the basis for the retrieval of LST through a radiative transfer equation via the single-channel method or more typically, the multi-channel method (Li et al., 2013). This study features a 30 year dataset (1979-2009) of LST retrievals from the High-Resolution-Infrared-Radiation-Sounder (HIRS) sensors that were flown on operational National-Oceanic-Atmospheric-Administration (NOAA) polar satellites (Shi and Bates, 2011). HIRS provides global LST retrievals potentially twice a day under clear sky conditions at a spatial resolution of 0.5° (Coccia et al., 2015). Due to HIRS's multi-decadal data record length it has been selected by the Global Energy and Water Exchanges Project (GEWEX) Data and Analysis Project (GDAP) as the primary satellite data source for the development of GDAP's internally consistent datasets. Hence the HIRS dataset will most likely receive more attention in the future with large-scale LSM validation being one possible application. However, it is important to reflect on HIRS's usability as an adequate LSM validation target in terms of accuracy, spatial resolution and temporal frequency, which is addressed in section 3.1.

Satellite remote sensing data with good spatial coverage enables a comprehensive spatial validation of a LSM. This provides information on spatial deficiencies that can help to diagnose model errors, which may remain undetected using station based hydrological data in a traditional validation (Koch et al., 2015b). However there exists no formal framework for assessing spatial performance of a model in an optimal way so that the information on spatial patterns is fully taken into consideration. The demand for true spatial performance metrics that go beyond simple cell to cell comparisons was highlighted by Wealands et al. (2005) who suggested innovative performance metrics in a soil moisture validation case study. For the field of atmospheric science, Gilleland et al. (2009) summarized various spatial metrics and categorized them into feature based, neighborhood, scale separation and field deformation approaches. Additionally, Wolf et al. (2014) compared standard metrics with innovative metrics, such as neighborhood- and object-based metrics, to validate predicted precipitation fields. Besides these efforts, there are only a limited number of spatial validation studies of land surface variables that fully embrace the availability of satellite remote sensing data by means of true spatial performance metrics.

The main feature of this study is the application of two innovative spatial performance metrics that are suitable for a comprehensive spatial model validation. First, an Empirical Orthogonal Functions (EOF) analysis is conducted jointly on observed and simulated LST maps to assess the similarity between the two datasets. In spatial validation studies by Fang et al. (2015) and Koch et al. (2015b) the EOF analysis proved to be very beneficial and insightful as a diagnostic validation tool. Second, a connectivity analysis is applied on warm and cold LST clusters that are derived by truncation of the simulated and observed LST fields at specific thresholds. Connectivity is a common metric in hydrogeology (Renard and Allard, 2013), but only few studies have implemented the concept of connectivity to characterize spatial patterns of surface variables (Western et al., 2001). Grayson et al.

(2002) underlined the physical meaning of connectivity of soil moisture patterns as a mechanism of controlling runoff. Further, their study showed that the connectivity analysis captured more adequate spatial information than the more standard variogram analysis. Both metrics are bias insensitive and thus focus on the spatial patterns as such. This is of special importance in a multi-model pattern validation study, because the models might have individual biases which should not interfere with the pattern comparison. Nevertheless, the bias is an integral measure of a model validation and should therefore always be considered separately. Furthermore, both metrics require high spatial coverage to guarantee a meaningful analysis and therefore their application is constrained to time steps with a low influence of cloud cover.

Actual evapotranspiration (aET) is a fundamental variable in the hydrological cycle and it is highly heterogeneous in time and space (Stisen et al., 2008). At local scale aET may be accurately measured by an eddy covariance tower (Alfieri et al., 2011). However, at larger scales there are not sufficient ground observations to account for the distinct spatial variability of aET. Satellite products cannot directly retrieve aET without relying on modelling. Therefore other variables, such as LST, can be used as a proxy for spatially distributed aET information (Anderson et al., 2011; Karnieli et al., 2010). This study reflects on the coupling between aET and LST based on in-situ observations at eddy covariance towers (Fluxnet) and how this coupling is represented in the LSMs. Further we investigate if apparent errors in predicted LST can be related to errors in predicted aET. From a process viewpoint it is generally expected that a cool LST bias is linked to a high aET bias through an overemphasized evaporative cooling.

The LSMs that undergo a spatial validation in this study are taken from the second phase of the multi institutional North American Data Assimilation System (NLDAS-2) (Xia et al., 2012a; Xia et al.,

2012b). NLDAS-2 provides high quality atmospheric forcing data and multi-model output of hourly hydrological variables over the contiguous USA (CONUS) since 1979 at a spatial resolution of 0.125° (~14km). In previous NLDAS studies GOES-East (Geostationary Operational Environmental Satellite, GOES-8) LST retrievals have been utilized to validate LSMs (Mitchell et al., 2004; Wei et al., 2013; Xia et al., 2015b). However, these studies missed the full potential of the validation dataset by only focusing on simple cell to cell metrics like the bias or the spatial correlation coefficient. Further, these studies were conducted on a limited validation period of several years, compared to the 30 year HIRS LST dataset used in this study.

The core novelty of this study is the development and testing of innovative spatial performance metrics that can expand the current validation toolbox of the modelling community. The NLDAS-2 models are selected because of the exhaustive validation groundwork in preliminary studies in which this spatial validation study can be nested. The HIRS LST dataset is chosen as the validation target, because of its availability, multi-decadal data record length and its valuable spatial coverage.

The aims of this study are **(1)** to present a comprehensive land-surface temperature (LST) dataset with global coverage that allows for a long-term validation of LSMs against monthly LST dynamics, **(2)** to introduce two innovative spatial performance metrics that are suitable for a thorough bias insensitive validation of simulated LST patterns, **(3)** to investigate the applicability of the HIRS LST dataset and the spatial metrics in a validation of the NLDAS-2 LSMs and **(4)** to examine the coupling between actual evapotranspiration (aET) and LST and reflect on the usability of LST as a proxy for diagnosing model representation of the water balance.

3. Methods and data

3.1. High-Resolution-Infrared-Radiation-Sounder (HIRS) LST Dataset

Remotely sensed data used to retrieve land surface temperature (LST) have been provided by different platforms since the late 1970's. Among them is the High-Resolution-Infrared-Radiometric-Sounder (HIRS), flown on board the NOAA polar orbiting satellites (Shi and Bates, 2011). The HIRS instrument has flown on 11 different satellites and has provided multispectral data since July 1979. HIRS LST retrievals are in swath format and available for clear sky conditions at 0.5° (~55km) spatial resolution with two return times per day at varying equatorial passing times (Coccia et al., 2015). For a more detailed technical description of the HIRS instrument we refer to Robel (2009).) The cloud detection follows the procedure presented by Jackson et al. (2003), where HIRS channel 8 (11.1 μm) brightness temperature is compared spatially and temporally with an estimated clear-sky value. If the deviation in brightness temperature is too cold (below a threshold) the observation is rejected as cloudy. The inter-satellite calibration by Shi (2011) resulted in fairly consistent LST retrievals between the satellites. Nevertheless, Siemann et al. (2016) highlighted that small inter-satellite biases still exist by comparing HIRS LST with twelve Baseline-Surface-Radiation-Network (BSRN) stations (Ohmura et al., 1998). The daytime biases of the satellites are ~ 0.5°C, varying from -0.1°C to 0.88°C, while the nighttime biases are usually higher (~1.5°C) and the range spans from 0.1°C to 2.1°C between the satellites.

This study utilizes the 30-year record (1979 - 2009) of hourly HIRS LST data over CONUS to validate the spatial patterns of simulated LST from the three LSMs. As with any other satellite retrieved LST product, the HIRS data is limited to cloud free conditions and thus exhibits spatial gaps. This makes an instantaneous hourly observation over CONUS unusable for an analysis of spatial LST patterns and

first at monthly time scale HIRS provides a reasonable coverage over CONUS. However, some cells are poorly represented, because the monthly average is either based on very few observations or the average is biased due to too many nighttime observations, because nighttime observations are more inclined to be cloud free than daytime observations. Thus two loose constraints are introduced for each grid cell to ensure representativeness: (1) A minimum of four observations per month and (2) a daytime nighttime ratio that does not exceed one to four. A radiation threshold of 100 W/m^2 (based on the NLDAS-2 forcing data) is chosen to distinguish between daytime and nighttime hours. Taking all CONUS data from the eleven satellites into consideration and applying the two above mentioned constraints yields a fractional coverage of 0.6 and higher for most months (Figure 1). The best coverages (>0.95) are during late summer and autumn. Orbital drift is an acknowledged issue of the NOAA satellites (Jackson and Soden, 2007; Wylie et al., 2005) which causes a shift in equatorial crossing time over the lifespan of a satellite (e.g. up to 3.8 hours for NOAA-14). However, orbital drift does not affect the validation of the NLDAS-2 simulations, because only grids that are collocated in time and space with an hourly HIRS observation are extracted from the model output and used for validation.

3.2. NLDAS-2

This study uses LSM data from the second phase of the multi institutional North American Data Assimilation System (NLDAS-2) (Xia et al., 2012a; Xia et al., 2012b). NLDAS aims at constructing datasets of hydrological states and fluxes of high spatial and temporal quality based on the best available observations for application in coupled model initialization, drought monitoring, and understanding hydrologic variability. This study focuses on three of the four LSMs: Mosaic (Koster and Suarez, 1992), Noah (Ek et al., 2003) and the Variable Infiltration Capacity (VIC) model (Wood et

al., 1997) which all incorporate a full soil-vegetation-atmospheric-transfer (SVAT) scheme. In comparison to NLDAS-1 (Mitchell et al., 2004), NLDAS-2 improved the accuracy and the consistency of the atmospheric datasets, upgraded the code and parametrization of the LSMs, and extended the simulation period from 3 years to more than 30 years. The NLDAS-2 LSMs provide hourly data for all relevant hydrological fluxes and state variables at a resolution of 0.125° (~14km) across CONUS from 1979 to present. The LSMs underwent thorough validations against streamflow data (Xia et al., 2012a), station based soil moisture data (Xia et al., 2014) and station based evapotranspiration data (Xia et al., 2015a). Additionally, Noah was individually validated against station based soil temperature (Xia et al., 2013) and satellite derived (GOES-8) LST (Wei et al., 2013; Xia et al., 2015b). Mitchel et al. (2004) evaluated LST for the NLDAS-1 LSMs, utilizing station based data for assessing the diurnal cycle and satellite based (GOES-8) data for assessing the spatial patterns, but was limited by the short simulation record. The NLDAS-1 study linked some of the LST disparities between the LSMs with the observations to differences in aerodynamic conductance (Noah), ground heat flux (VIC) and canopy conductance (Mosaic). For NLDAS-2, Xia et al. (2012b) suggested that the differences between Noah's and Mosaic's spatial LST patterns over CONUS was explained by their differences in albedo. Areas of higher and lower albedo were clearly negatively correlated to differences in LST. The overall higher albedo in Noah caused lower net shortwave radiation, which corresponded well to the generally cooler LST in Noah compared to Mosaic. Despite these previous efforts to validate simulated LST in the NLDAS LSMs, a thorough spatial validation of the simulated patterns has not been conducted yet. Previous studies applied simple cell to cell metrics and thereby lacked a true pattern comparison. Furthermore, the 30-year coverage of HIRS allows a LST validation of the entire NLDAS-2 simulation period, which has not been undertaken yet. Further the NLDAS LSM output is resampled from its original 0.125° resolution to 0.5° to provide consistency with the HIRS LST data.

3.3. FLUXNET

Fluxnet is a global network of micrometeorological flux measurement stations (Baldocchi et al., 2001) that provides high quality data on water-, energy- and carbon-fluxes across a diverse range of ecosystems and climates for multiple years. This study uses data from 74 stations that are located across the U.S. and are part of the American AmeriFlux network. Flux data are measured half-hourly from 1991 to 2007, but not all stations cover the entire period nor have complete measurements of all fluxes. Moreover, the flux data is screened for energy balance closure at monthly time scale following the approach presented by Stoy et al. (2013) and Wilson et al. (2002) The quality controlled data are used for two purposes in this study: (1) To spatially and temporally validate the HIRS LST observations with in-situ data and (2) to explore the coupling between HIRS LST and in-situ actual evapotranspiration (aET), and investigate if the LSMs exhibit a comparable coupling. It has to be noted that the differences in the spatial footprint and scale complicate a comparison between in-situ data from flux towers and large scale satellite data and can cause inconsistency in the validation of satellite data (McCabe and Wood, 2006). The effect of the mismatch in spatial footprint is not directly quantified in this study. Instead the diurnal variability of satellite and in-situ LST is assessed at three Fluxnet sites to facilitate a better understanding of the differences in the diurnal signal due to the differences in scale. The three sites are situated in distinctly different climates and are selected as examples to discuss the diurnal variability of HIRS and the general effect of scale mismatch.

Many studies have utilized the measured surface longwave radiation data at the Fluxnet stations to validate remotely sensed LST products (Cleugh et al., 2007; Trigo et al., 2008; Wang and Liang, 2009). LST can be related to surface longwave radiation by the Stefan Boltzmann law and reformulated as:

$$LST = L\uparrow - 1 - \varepsilon \cdot L\downarrow \varepsilon \cdot \sigma 14 \qquad \text{eq.1}$$

where $L\uparrow$ and $L\downarrow$ are the upward and downward longwave radiation, ϵ is the surface emissivity and σ is the Stefan–Boltzmann's constant $(5.67 \cdot 10{-}8\ W\ m{-}2K{-}4)$. As the HIRS LST retrievals assume a constant surface emissivity of 1, the apparent relationship in equation 1 is purely driven by the upward longwave radiation. 15 Fluxnet stations across CONUS feature longwave radiation measurements and monthly LST averages are only used in the subsequent analysis if 90% of the half-hourly $L\uparrow$ data are available in the respective month.

The eddy covariance data at the Fluxnet sites has frequently been incorporated in various studies to derive in-situ aET observations (Cleugh et al., 2007; Jung et al., 2010; Velpuri et al., 2013). Following Mu et al. (2011), aET in terms of water depth can be derived from the latent heat flux (*LE*) measured at the Fluxnet eddy covariance stations:

$$aET = LE\lambda \qquad \text{eq. 2}$$

where λ is the latent heat of vaporization $(J\ kg{-}1)$ that depends on the air Temperature Ta. The *LE* data is measured half-hourly at the eddy covariance towers and each 30 min aET (mm) is calculated as:

$$\lambda = 2.501 - 0.002361 \cdot Ta \cdot 106 \qquad \text{eq. 3}$$

$$aET = LE \cdot 60 \cdot 30\lambda \qquad \text{eq. 4}$$

Monthly averages of aET are only used at 51 stations with eddy covariance data for months with at least 90% of measurements of half-hourly *LE* and T_a.

3.4. Spatial performance metrics

This study features two innovative spatial performance metrics that enable a meaningful quantitative validation of simulated LST spatial patterns. The metrics are derived from (1) an EOF analysis and (2) a connectivity analysis of the simulated and observed LST patterns. Both metrics are bias insensitive, which is favorable for this multi-model spatial validation, because individual model biases might interfere with the validation. Furthermore, these metrics require good spatial coverage in order to produce meaningful results. This is especially the case for the connectivity analysis which is therefore only conducted on months with a coverage greater than 0.95. On the other hand, full coverage is less essential for the EOF analysis where the coverage threshold is set to 0.9. This constrains the spatial validation to 33 and 91 months out of 30 years, respectively.

3.4.1. EOF analysis

The Empirical-Orthogonal-Functions (EOF) analysis is a frequently applied statistical methodology in the hydrological community to assess large spatio-temporal datasets of hydrological states and fluxes. Most commonly it has been applied to observed (Korres et al., 2010; Perry and Niemann, 2007) soil moisture data, but a recent application highlighted its applicability to surface fluxes as well (Mascaro et al., 2015). The main feature of the EOF analysis is that it decomposes the variability of a spatio-temporal dataset into a set of orthogonal spatial patterns (EOFs) that are invariant in time and a set of loadings that describe how the EOFs are weighted over time. The spatial pattern of the first EOF always captures as much as possible of the variance and the following EOFs will subsequently add to the explained variance. For a detailed description of the methodology we refer to Graf et al. (2014). The EOF analysis is typically applied on observational or modeled datasets to understand spatio-temporal variability, however recent applications stressed it's usability as a tool for a comprehensive

spatial validation of distributed hydrological models at catchment scale (Fang et al., 2015; Koch et al., 2015b). In order to derive a quantitative spatial performance metric Koch et al. (2015b) suggested to conduct a joint EOF analysis on both observed and simulated data. In this way, the resulting EOF maps honor the spatio-temporal variability of both datasets and the weighted difference in the loadings at specific times can be utilized to derive a meaningful pattern similarity score. The weighting is required, because each EOF contributes differently to the explained variance. Thus the EOF based similarity score (S_{EOF}) between an observed and a predicted LST map at time x can be formulated as:

$$SEOFx=i=1nwiloadisimx-loadiobsx \quad \text{eq. 5}$$

where w_i, the variance contribution of the i'th EOF, is multiplied with the absolute difference between the simulated loading ($load^{sim}$) and the observed loading ($load^{obs}$) of the i'th EOF at time x. Prior to the EOF analysis the monthly mean is removed from each LST map, thus the methodology is based on the spatial anomalies which makes it bias insensitive.

3.4.2. Connectivity analysis

Within the field of hydrogeology, connectivity is a widespread measure to characterize the heterogeneity of an aquifer (dell Arciprete et al., 2012; Koch et al., 2014). From a hydrogeological perspective, the degree of connectivity has direct physical implications on groundwater flow and solute transport. Western et al. (2001) and Grayson et al. (2002) are among the few studies that applied a connectivity analysis on land surface variables. Both studies analyze soil moisture patterns at a small catchment in Australia (Tarrawarra, 10.5 ha) and were able to link soil moisture connectivity to runoff behavior. This finding also stresses the physical relevance of connectivity as a characteristic of spatial patterns of other hydrological states such as LST. Another typical application that incorporates the

concept of the connectivity of hydrological variables is the identification and tracking of drought events (Andreadis et al., 2005; Sheffield et al., 2009).

On a regular grid the connectivity of a binary variable can either be via faces or via corners; both having four possible connections. Connectivity via faces comprises cells that are vertically and horizontally adjacent whereas connectivity via corners describes the diagonal direction. In this study we consider both of them which results in eight possible connectors per cell. Furthermore, two cells are connected if there exists a sequence of neighboring cells between them. Connected cells can then be grouped into individual clusters. In order to apply this methodology on continuous variables, such as LST or soil moisture, Renard and Allard (2013) suggested to decompose the continuous field, denoted as $Y(x)$, into a series of binary sets. The simplest way to decompose $Y(x)$ is to introduce an increasing threshold t which stepwise, from minimum to maximum, truncates the field into a binary variable X:

$$Xt=x:Yx\geq t. \qquad \text{eq. 6}$$

This will generate a series of binary cluster maps where $Xt1\subset Xt2$ if $t1>t2$. In the case of LST, t classifies the continuous LST field into a binary map of cold and warm clusters. For this study, the threshold value moves along all percentiles of the LST range and generates a series of 100 binary maps of cold and warm clusters. Focusing on the percentiles makes this methodology bias insensitive and in fact it allows to compare the spatial patterns of two different variables that are expected to be correlated (e.g. soil moisture and actual evapotranspiration). Next, percolation theory can be used to describe the transition from many disconnected clusters to a very large spanning cluster as t increases. Hovadik and Larue (2007) suggested the probability of connection as a suitable metric to quantify how percolated

clusters are. The metric, denoted as $\Gamma(t)$, is computed for each threshold (t) as the proportion of the pairs of cells that are connected among all possible pairs of connected cells:

$$\Gamma t=1nt2i=1NXtni2 \quad \text{eq. 7}$$

n_t is the total number of cells in the binary map X_t at threshold percentile t. n_i is the number of cells in the i'th cluster in X_t which has $N(X_t)$ distinct clusters in total. Renard and Allard (2013) plotted the resulting connectivity curves, $\Gamma(t)$, for different synthetic fields, and underlined that patterns are equipped with a unique connectivity curve. Especially for the percolation threshold, the specific threshold at which the connectivity abruptly increases is a very distinct characteristic for each pattern. Based on numerical tests on synthetic 2D rectangular domains, Hovadik and Larue (2007) estimated the percolation threshold to be 0.59 for a four-edge-connectivity. With regard to LST patterns, which are underlain by an intrinsic autocorrelation and are thus not placed randomly in space, the percolation threshold is expected to be generally lower. Overall, $\Gamma(t)$ can be understood as a measure of homogeneity and smoothness of the patterns. A major benefit of the connectivity analysis is that it allows for a separate assessment of cold patterns (cold phase: From coldest to warmest percentile) and warm patterns (warm phase: From warmest to coldest percentile). Grayson et al. (2002) applied the connectivity function in their analysis of soil moisture patterns, which is a more elaborated connectivity analysis than the probability of connection, $\Gamma(t)$, used in this study. The connectivity function reflects the probability that a cell of a binary map is connected with another cell (i.e. both are in the same cluster) as a function of distance. The Grayson et al. (2002) study highlighted that the connectivity function, as a way to characterize spatial variability, can contain more spatial information than the more common variogram analysis. Renard and Allard (2013) identify a relationship between the sum of

335 the connectivity function and $\Gamma(t)$, which supports the use of $\Gamma(t)$ which clearly is the simpler metric to
336 compute and interpret.

337 In order to derive a quantitative measure of how good the observed LST connectivity ($\Gamma(t)obs$) is
338 represented by a model ($\Gamma(t)sim$), the root-mean-squared-error ($RMSE_{Con}$) between the observed and
339 simulated connectivity curves, $\Gamma(t)$, can be computed for both phases:

340 $$RMSECon=t=1100\Gamma(t)obs-\Gamma(t)sim2100 \quad \text{eq. 8}$$

341 In this context, the RMSE provides a global skill assessment of the connectivity that is not constrained
342 by local agreement. Hence, the structure of the patterns may match, but the corrected allocation of the
343 patterns is not warranted. This in analogous to the comparison of observed and predicted
344 semivariograms (Korres et al., 2015).

345 # 4. Results and discussion

346 ## 4.1. HIRS

347 Before addressing the validation of HIRS against in-situ LST data at Fluxnet sites we want to broadly
348 discuss the usability of HIRS as a validation target, put HIRS into perspective to other satellite LST
349 products and reflect HIRS's spatial and temporal limitations.

350 In general, polar orbiting satellites allow an insightful analysis of spatial processes, but their low
351 overpass frequency limits an adequate temporal analysis. In contrast, geostationary satellites can fill the
352 temporal gap and provide high resolution temporal data on diurnal processes, but are equipped with a
353 fixed viewing window that hinders global coverage. In theory, various geostationary satellites could be

mosaicked in space and time to a global product, which, to our knowledge, has not been attempted yet. Ideally a combination of both should be considered for a holistic validation of land-surface processes which are complex in time and space. However, the incorporation of both polar orbiting and geostationary LST retrievals in a single validation is beyond the scope of this study. Gunshor et al. (2004) underlined that the calibrated infrared brightness temperature retrieved by polar orbiting satellites (HIRS and AVHHR: Advanced Very High Resolution Radiometer) and geostationary satellites (GOES-8, -10 and Meteosat-5, -7) and concluded that all instruments show small differences within 0.6 °C. Despite HIRS's accuracy, which is in reasonable agreement with other sensors, there are issues concerning the spatial and temporal resolution of the retrievals. The 0.5° spatial resolution of HIRS is coarser than alternative polar orbiting satellites such as AVHHR (Frey et al., 2012; Heidinger et al., 2013) or MODIS (Wan et al., 2002; Wan et al., 2004). Nonetheless it still provides valuable spatial information for the assessment of continental to global scale LSMs. Current global assessments of water budgets (e.g. Rodell et al., 2015) and state-of-the-art LSMs and hydrological models (e.g. Haddeland et al., 2011) are at resolutions of the order 0.5°, commensurate with the HIRS data. The native resolution of continental LSMs might be finer (e.g. NLDAS-2), but the predictive capability at the fine scale is questionable, given inadequate parameterizations and meteorological observations in many parts of the world, particularly for precipitation (Sheffield et al., 2014); thus an aggregation to 0.5° reduces uncertainty and seems reasonable if the predominant spatial patterns across a continent are of interest. Compared to other LST products, HIRS assumes a constant surface emissivity of one, which makes it a favorable validation dataset, because the same assumption is most commonly applied in LSM applications (Mitchell et al., 2004).

Other LST products may provide more detail in time or space, but HIRS can still be regarded as a valuable observation if large scale LST patterns over a multi-decadal period are of interest. Additionally, the 30-year dataset was first processed at the National Climatic Data Center (Shi, 2011) and very recently applied by Coccia et al. (2015) and Siemann et al. (2016) to generate a global hourly LST dataset using a Bayesian merging procedure that combines HIRS with reanalysis LST data. This study wants to expand the applicability of this recently introduced LST dataset by exploring the usability of HIRS for the spatial validation of LSMs.

In order to ensure the accuracy of the HIRS LST dataset over CONUS, this study first conducts a validation of the remote sensing observations against in-situ observations at Fluxnet sites. First, monthly values are accessed at the flux sites and compared with monthly averages of collocated HIRS observations. Secondly, the diurnal variability of HIRS LST is addressed at three Fluxnet sites for July 2004, where four NOAA satellites measured simultaneously which gives eight potential overpasses a day. Figure 2 depicts the results based on 511 monthly LST averages at 15 stations that measure upward longwave radiation (2000-2006). In spite of differences in the temporal coverage between stations, Figure 2 does not distinguish between different years and analyzes the entire datasets jointly. The scatter plot (Figure 2 b)) reveals a strong temporal correlation between in-situ LST and satellite retrieved LST alongside a warm bias of the HIRS data of 1.9°C. Figure 2 a) disaggregates the scatter plot into individual stations and plots their biases on a CONUS map. All stations exhibit a strong temporal correlation of > 0.95 and are generally characterized by a warm bias, besides two stations that have a cold bias. The combined root-mean-squared-error (RMSE) between HIRS LST and Fluxnet LST is 3.7°C and the individual RMSE per station lies between 1.7°C to 8.8°C. In order to validate HIRS across seasons and across climate zones, the spatial correlation coefficient is computed for each month

from 2000 to 2006 (not shown). Each month is covered by LST data from at least 9 Fluxnet stations and the average spatial correlation is 0.84. Further, only six months out of the seven years show a spatial correlation below 0.7. Siemann et al. (2016) conducted a global validation of the hourly HIRS observations against the Baseline-Surface-Radiation-Network (BSRN) (Ohmura et al., 1998). 7 out of the 12 BSRN sites are situated in CONUS and the overall correlation with the HIRS LST retrievals is comparable to the Fluxnet correlations. The validation in Siemann et al. (2016) was based on hourly data and split up into daytime and nighttime. In both cases HIRS manifests a warm bias, but the nighttime basis is generally higher (~ 1.5°C) than the daytime bias (~ 0.5°C). In summary, the Fluxnet validation is comparable to the BSRN validation and reassures the accuracy of the HIRS LST dataset and thus its reliability for a spatial model validation. However, for further applications of the HIRS LST dataset it is important to be aware of its warm bias. There is only limited information on the spatial structure of the bias and therefore it is not taken into account during the spatial validation in this study. Figure 3 addresses the diurnal variability of HIRS LST at three Fluxnet stations that are situated in distinctly different climate conditions across the U.S. and mean monthly values are given for each hour in July 2004. In that period, four NOAA (14 - 17) satellites operated simultaneously which supplies eight potential overpasses a day. The diurnal amplitude of HIRS LST seems reasonable in comparison to the Fluxnet data the at three given sites. However the previously discussed warm bias is clearly visible but differs temporally between the sites: The Montana site has a pronounced midday warm bias, the Illinois site shows a rather constant warm bias over the entire day and lastly, at the Arizona site the nighttime warm bias is most emphasized. This complex spatio-temporal behavior of the HIRS bias is expected to be caused by differences in spatial footprint between the in-situ data and the satellite retrievals. The tracks of the NOAA satellites vary from day to day, while remaining the same equatorial crossing time, thus the observation time shifts between days for the two overpasses that

420 are recorded for each satellite. The uneven distribution of observations shown in figure 3 emphasizes
421 the limited applicability of HIRS for the validation of diurnal processes and instead geostationary
422 products such as GOES-8 would clearly be more suitable for a task like this.

423 **4.2. Spatial validation of LST patterns**

424 The overall goal of this study is to conduct a comprehensive spatial model validation of the NLDAS
425 LSMs using innovative performance metrics. Before applying these, a more general assessment of the
426 spatial performance is presented in the following section.

427 In general, remotely sensed LST (HIRS) and simulated LST (NLDAS) are both related to an
428 instantaneous radiometric surface temperature based on the upwelling longwave surface radiation and
429 are therefore comparable. In order to facilitate a fair comparison, only spatially and temporally
430 collocated hourly LST data are extracted from the LSMs at grids where HIRS provides a cloud free
431 observation for computing the average monthly LST maps. All LST data incorporated in this study, (1)
432 the NLDAS LSMs, (2) the in-situ Fluxnet sites and (3) the HIRS retrievals, underlie the assumption of
433 a constant surface emissivity of one. This underrepresentation of heterogeneity in time and space may
434 introduce errors, but at the same time it may also cancel out, because the same assumption is applied to
435 all datasets. This assumption is in general most valid for dense vegetation or snow, but less applicable
436 for bare soils.

437 Figure 4 presents monthly HIRS LST maps of two example months (March and August, 2004) and the
438 average LST map based on all monthly data in the 30 year period (1979-2009). The seasonality in the
439 observed LST data is striking, as the patterns drastically change from a cold month (March 2004) to a
440 warm month (August 2004). Figure 4 also features the bias maps of the three LSMs for the respective

observations. All LSMs display seasonality in their bias maps, hence areas with a warm bias change to a cold bias between the two months or vice versa. The common features among the LSMs bias maps are the warm bias in the northeast in March 2004 and the warm bias in Texas for both months. VIC generally has the most complex seasonality whereas Mosaic and Noah reflect a rather constant cold bias over entire CONUS throughout the months. The similarity between Mosaic's and Noah's LST patterns and the dissimilarities between them and VIC have already been pointed out by Xia et al. (2012b). Xia et al. (2015b) validated Noah against GOES-8 nighttime LST over CONUS for a 13 year period (1997-2009) and the magnitude and the pattern of the bias map resemble the one presented in Figure 3.

Figure 5 focuses on the temporal component of the LST validation. The top panel depicts the monthly mean LST anomaly for the observations (HIRS) and the three LSMs. While all datasets have a distinct seasonality and all reflect some inter-annual variability, VIC clearly has the lowest amplitude with too warm winters and too cold summers. This is supported by the bottom panel of Figure 5, which shows the bias and spatial correlation per month. Mosaic and Noah have a uniform cool bias of ~-3°C, while VIC has a distinct seasonality in its bias with a warm bias in winter (~3°C) opposed to a slight cool bias in summer (~-1°C). The biases of the LSMs are clearly elevated in the first few years (1979-1984). To our knowledge, no inter-satellite validation of the HIRS LST data has been conducted for the early period (NOAA-06, 07 and 08). Small inter-satellite biases can be derived from the work by Siemann et al. (2016) from NOAA-11 and onwards. Thus it can only be speculated if the elevated LSM bias in the first few years is related to biases in HIRS LST or to biases in the NLDAS forcing in that period. The spatial correlation coefficient of the simulated and observed monthly LST maps is generally very good (>0.8) and the LSMs show a similar behavior apart from VIC which shows single low correlation

outliers in the month of January for some years. Figure 6 summarizes the results described above by showing the monthly averages of the bias and the spatial correlation coefficient for the 30 years of validation. The distinct seasonality of the VIC bias is very apparent opposed to the rather constant cold bias of Mosaic and Noah. The average correlation coefficient has small seasonality and it is generally very satisfying for all 3 LSMs.

Plotting the mean versus the standard deviation is widely used to assess the spatio-temporal variability of soil moisture patterns (Famiglietti et al., 2008; Graf et al., 2014). In the case of soil moisture the relationship is typically defined by an upward convex behavior with highest spatial variability at very wet or very dry conditions. Figure 7 presents the results for the monthly observed (HIRS) and simulated LST data. The HIRS LST data clearly reveal a linear relationship between the monthly mean LST and its spatial variability with higher variability in colder months. This is caused by the distinct climate variability over CONUS, which is characterized by homogeneously warm LST patterns during summer months and an enhanced LST variability during winter months due to a distinct separation of a warm south and a cold north. Mosaic and Noah exhibit a similar relationship, although their cold bias can clearly be detected, as all months are slightly shifted towards colder LST. VIC follows the observed linear relationship until ~5°C and for lower temperatures the spatial variability drops. Besides the lack of spatial variability the warm bias of VIC is also noticeable as the cold months are shifted towards warmer LST.

4.2.1. EOF analysis

The previous analysis revealed that VIC has the most complex LST deficiencies with a clear seasonal signal in the bias and too little spatial variability during cold months. Therefore the results of the EOF

484 analysis are discussed for VIC in more detail and the results for Mosaic and Noah are briefly
485 summarized at a later stage.

486 Due to prior mean removal, the EOF analysis is a bias insensitive approach and thus it is not affected
487 by the bias seasonality shown in Figure 6. A joint EOF analysis is conducted for both observed (HIRS)
488 and simulated (VIC) monthly LST maps for 91 months that have a spatial coverage greater than 0.9.
489 The EOF maps in Figure 8 represent the predominant spatial patterns that are found in the 182
490 observed and simulated LST maps. The first EOF can capture 76% of the total variance and expresses
491 the most underlying pattern of the general warm-cold LST gradient from South to North. Additionally
492 high altitude areas in the western mountains are identified with the lowest values. Generally, the values
493 of the EOF maps do not have a direct physical meaning as such. First, when an EOF map is multiplied
494 with its loadings, the resulting product can be understood as a deviation in °C from the mean. The
495 pattern of the second EOF, which contributes additional 6% to the explained variance, is more complex
496 and its physical meaning is first revealed after assessing its loadings. The subsequent EOFs express less
497 than 2% of the variance and therefore they can be considered as noise originating either from the HIRS
498 observations or the LSMs. The loadings are presented in Figure 8 and the sign of the loadings for the
499 second EOF switches from positive in summer to negative in winter, which results in a seasonal
500 inversion of the pattern. For example, the Great Plains (positive EOF2 values) are “extra” hot in
501 summer and “extra” cold in winter whereas many of the coastal areas (negative EOF2 values) have
502 “milder” LST with warmer winters and colder summers. Comparing the loadings of the observed and
503 simulated LST maps in Figure 9 reveals that the second EOF is better represented by VIC than the first
504 EOF. Similar to the mean versus standard deviation plot in Figure 7, the loadings of VIC for the first
505 EOF are too low during colder months. This translates to too small spatial variability during those

506 months, because the predominant South-North gradient in EOF1 is weighted too little by VIC. The
507 EOF based similarity is derived from the weighted sum of the differences in loadings between HIRS
508 and VIC in equation 5 and, based on Figure 9, it can already be anticipated that poor performance is
509 attested to the cold months.

510 The resulting EOF maps for the validation of Mosaic and Noah are almost identical to the ones of VIC
511 in Figure 8 and therefore not shown. On the other hand, the derived EOF based similarity scores for the
512 three LSMs are different as presented in Figure 10. The EOF based metric rates the spatial performance
513 of Mosaic and Noah as very similar for the warmer months and attests diverging similarities to the two
514 LSMs for the colder months. Following the EOF analysis, the LST patterns in the warmer months are
515 explicitly better predicted by VIC than by Mosaic and Noah. On the contrary, Mosaic and Noah clearly
516 provide a better spatial performance than VIC for the colder months.

517 Several studies (Jawson and Niemann, 2007; Qiu et al., 2014) tried to identify the main drivers of
518 spatial variability of soil moisture by conducting an EOF analysis and subsequently calculating the
519 spatial correlation between the resulting EOF maps of soil moisture with EOF maps of potential drivers
520 (e.g. precipitation, topography, vegetation). Important drivers were identifiable by a strong correlation.
521 For the LST case, the pattern of the first EOF in Figure 8 correlates strongly (0.86) to the first EOF of
522 air temperature, which emphasizes the strong physical coupling between the atmosphere and surface,
523 which VIC captures better in the warm months than in the cold months.

524 **4.2.2. Connectivity analysis**

525 Following the description in section 3.4.2., the simulated and observed LST maps can be assessed and
526 quantitatively compared by means of a connectivity analysis. Each percentile of the temperature range

527 is utilized to generate a binary map of cold and warm which then undergoes a cluster analysis. Figure
528 11 exemplifies the cluster analysis of observed (HIRS) and simulated (VIC) LST for August 1993 for
529 four different thresholds: 5^{th}, 20^{th}, 80^{th} and 95^{th} percentile. The thresholds correspond to the coldest 5%,
530 coldest 20%, warmest 20% and warmest 5%, respectively. Each distinct cluster is displayed with a
531 unique color and a first visual inspection indicates resemblance in location, size and number of clusters
532 between HIRS and VIC. For a complete and systematic analysis of the cluster maps at all percentiles,
533 the probability of connection is introduced as a metric. Figure 12 depicts $\Gamma(t)$, the probability of
534 connection, as a function of the threshold value t for the warm and cold phase of the observed and
535 simulated LST patterns presented in Figure 11. The LST patterns display an inherent autocorrelation,
536 therefore the connectivity is already high at very low percentiles. As the threshold value increases for
537 the warm and cold phase, connectivity generally increases as well. The connectivity curves of the
538 observed LST have unique shapes with distinct percolation thresholds where the probability of
539 connection increases abruptly. The LSMs generally reflect the percolation thresholds in position and
540 magnitude quite well and the three LSMs are overall very similar in terms of their LST connectivity.
541 The most apparent difference between the LSMs is that VIC's warm phase clearly percolates earlier
542 than Mosaic and Noah. This can be attributed to a larger degree of homogeneity in VIC's warm
543 patterns. The RMSE between the connectivity curves of HIRS and the LSMs (eq.8) can be used as a
544 quantitative metric to assess the spatial performance of the warm and cold phase separately for each
545 LSM.

546 In total, 33 months of high coverage (>0.95) are incorporated for the connectivity analysis. Most of
547 them are in August (11), September (12) and October (8). Figure 13 illustrates the average connectivity
548 curves for the three months derived from the HIRS data and from the three LSMs. This allows a

549 detailed analysis of the evolution of the LST patterns during the transition from summer to winter. The
550 observed connectivity curves clearly become steeper, when moving from August to October, and show
551 earlier percolations. Hence, cold months exhibit a more distinct separation between cold and warm
552 areas in comparison to warm months. In August the LST gradient in the HIRS data is smaller and the
553 transition from cold to warm is rather discontinuous and heterogeneous. On the other hand, the LST
554 range in October is expected to be larger and the clear north south gradient is more pronounced than in
555 August (EOF1). The continuous transition in October results in the steeper connectivity curves of the
556 cold and warm phase for the HIRS data in Figure 13. Generally, the LSMs behave quite similar in
557 terms of their connectivity and it is difficult to point out a single LSM with the best performance;
558 however the inter LSM similarity is more distinct for the warm phase than for the cold phase. The best
559 performance can be assessed for all LSMs for the warm phase in September and October. In those
560 months the warm patterns are simpler to model as they are mostly constrained to the southern part of
561 CONUS. Whereas in August, the warm patterns are more complex, because warm areas are less
562 localized and the LSMs do not capture the complexity in the patterns correctly and thus overestimate
563 the connectivity. The interpretation for the simulated connectivity of the cold phase in August is
564 analogous. Moving from August to October, the agreement between the connectivity of the cold phase
565 between HIRS and the LSMs declines, which is opposite to the warm phase. The connectivity of the
566 cold patterns in October is underestimated by the LSMs, meaning that the patterns are too
567 heterogeneous with respect to the observations.

568 **4.2.3. Comparison of metrics**

569 This study features two innovative spatial performance metrics that clearly require more effort to
570 implement compared to simpler cell to cell comparisons; such as RMSE or spatial correlation

coefficient (R). The clear advantage of both the EOF analysis and the connectivity analysis, over cell to cell comparisons, is that they offer additional features to the purely quantitative skill score. For instance, the EOF analysis provides EOF maps that represent the predominant spatial patterns and the connectivity analysis can be interpreted separately for cold and warm patterns. Both features provide rather qualitative insights for the spatial validation. Nevertheless, if applied in an automated calibration the qualitative features have no merit and only a single number, quantifying the spatial performance of the model, is of interest. Therefore, we analyze if the EOF and connectivity analysis hold additional information in comparison to more standard and simpler metrics like RMSE or R or if their information is redundant.

Figure 14 depicts the resulting performance metrics for VIC derived from the EOF analysis, RMSE and R for the 91 months used for the EOF analysis. Additionally, the performance derived from the connectivity analysis is given for the 33 months with coverage greater than 0.95. The warm phase is generally rated with a better performance than the cold phase and Table 1 underlines that the warm phase has noteworthy correlations with the RMSE (0.6) and R (-0.5). Strong correlations between two metrics indicate that their information content can be regarded as redundant. Figure 13 stresses that this is the most evident for the RMSE and EOF analysis in VIC, which has a correlation of 0.8. In the case of VIC the information provided by the EOF analysis and the connectivity analysis of the warm phase is partly already represented by the RMSE and R. The connectivity analysis of the cold phase shows no significant correlations to any other performance criteria in any of the LSMs (Table1). All metrics compared in Table 1 are meaningful, thus a metric with purely weak correlations to all other metrics does not imply that it is not informative, it rather implies that it contains additional information on the pattern performance compared to the other metrics. The correlations between the metrics are different

593 between the LSMs, but Mosaic and Noah have similar correlations. Taking the bias maps in Figure 3
594 into consideration underlines that the spatial pattern of the biases of Mosaic and Noah are similar and
595 VIC exhibits a very different pattern in its spatial bias. This indicates that the type of spatial error
596 controls whether two metrics provide redundant information or not; e.g. RMSE and EOF are strongly
597 correlated in VIC but have a weak correlation in Mosaic and Noah. This complicates the choice of
598 spatial performance metric, because metrics show no unique correlations to other metrics and their
599 sensitivity depends on the kind of spatial error that is evident. The EOF analysis as well as the
600 connectivity analysis is constrained to months with a high spatial coverage and Figure 14 gives the
601 distribution of months that fulfill the coverage criterion for the given metrics. Coverage is generally
602 highest in spring and autumn, but all months are included in the analysis although they are unevenly
603 represented.

604 **4.3. aET – LST coupling**

605 The previous section describes the results of the spatial model validation and underlines that the EOF
606 analysis and connectivity analysis reveals comprehensive insight into the LST related model
607 deficiencies. This section reflects on the implications of LST errors in LSMs for the energy and water
608 balance to guide the interpretation of spatial LST deficiencies. In this context we analyze actual ET
609 (aET) measurements at the Fluxnet sites. aET links the water and energy balance and from a process
610 viewpoint (evaporative cooling) it can be expected that an overestimation in aET is associated with a
611 cool bias in LST and vice versa. If this relationship is tangible, LST can be theoretically used as a
612 proxy to indirectly validate the spatial distribution of the water balance via aET. This is otherwise not
613 feasible because no components of the water balance are observable directly via remote sensing. On the

614 other hand, flux towers provide good temporal coverage, but their low spatial density and small support
615 scale limits the usability of tower data for a spatial validation of the water balance.

616 Figure 14 validates simulated monthly aET at the Fluxnet sites for VIC. The scatter plot in Figure 15 b)
617 reveals a negative bias of -5.3mm per month based on 2300 months at 51 Fluxnet sites over CONUS.
618 The correlation of 0.77 is reasonable, but the scatterplot identifies single months with very large errors
619 (>100mm/month). The overall temporal correlation and bias for Mosaic and Noah are 0.72,
620 17.5mm/month and 0.80, and -6.8mm/month, respectively. In general, the large positive aET bias for
621 Mosaic corresponds well with the cool LST bias. However the negative aET bias for Noah and its
622 generally cool LST bias contradict the expected relationship. The map over CONUS (Figure 15 a))
623 displays the VIC aET biases per station and the heterogeneous spatial pattern of positive and negative
624 biases stresses that there is no systematic spatial aET bias. The temporal correlations at the individual
625 stations are 0.83 on average with a minimum of 0.56.

626 It remains unclear how the aET errors in Figure 14 are related to errors in LST. In order to understand
627 the coupling between LST and aET better, Figure 16 investigates the relationship between the
628 hydrological state variable LST and the flux variable aET in more detail. This analysis is constrained to
629 daytime LST only, because it is expected that daytime LST is closer related to aET than nighttime LST.
630 Ideally the daily LST amplitude (daytime - nighttime) should be used to assess the link between LST
631 and aET, but due to the irregular distribution of HIRS overpass times it is not possible to compute a
632 meaningful LST amplitude based on the HIRS data. The observational data is based on monthly HIRS
633 daytime LST and Fluxnet aET at the 51 sites given in Figure 15. The coupling between the two
634 variables is of exponential nature with rapidly increasing aET as monthly daytime LST increases.
635 Another interesting feature is that the spread in monthly aET increases as well, because some of the

636 data express water limitation while others are characterized by energy limitation. Water limitation is
637 identified by high daytime LST and low aET and for energy limited conditions the aET can increase
638 exponentially alongside an increase in LST. The three models feature this increase in aET variability
639 for warmer months accordingly, but the general relationship between aET and daytime LST varies
640 between the LSMs. The fitted curves of Mosaic and Noah are shifted towards cooler LST, because of
641 their inherent cold bias. Mosaic clearly overpredicts aET across the entire daytime LST range, while
642 Noah is in good agreement with the observations. VIC's warm bias during cold months is clearly
643 apparent and it is consistent with an underestimation of aET. Out of the three LSMs, VIC seems to be
644 best at capturing the water and energy limiting control for the warmest months. Overall, the aET-LST
645 coupling is best represented by Noah and deviations in Mosaic and VIC are comprehensible from a
646 process viewpoint; too high aET is associated with cooling and vice versa.

647 **4.4. Diagnosis of spatial model errors**

648 LST is an important yet complex hydrological state variable of land-atmosphere interactions. The EOF
649 analysis identifies the strong coupling to air temperature and the previous section highlights the
650 complex relationship to aET. The comprehensive spatial validation of simulated LST patterns is
651 insightful and can be used as a diagnostic tool to learn about a LSM. However we have not touched
652 upon potential causes of the spatial deficiencies that are highlighted by the EOF and the connectivity
653 analysis. Attributing the problem to a general cause is rarely possible as a short literature review on
654 LST validation studies reveals. Wang et al. (2009) found that air temperature, especially the
655 temperature gradient for high altitudes, was a main concern in their LST validation. Koch et al. (2015b)
656 identified an overemphasized groundwater coupling, which resulted in a distinct cool LST bias as a
657 major limitation to their LSM. Silvestro et al. (2013) mentioned soil moisture and its effect on the

658 thermal inertia as a drawback in their LST predictions. Wei et al. (2013) relied on the parametrization
659 of vegetation (e.g. spatio-temporal variation of LAI, root density and stomatal resistance) to improve
660 LST predictions of a LSM. Lastly, Mitchell et al. (2004) focused on improving the energy fluxes in the
661 NLDAS-1 simulations by means of adjusting the aerodynamic conductance and the ground heat storage
662 term to get better LST estimates. Some of these issues focus on the diurnal and others on the annual
663 cycle of LST, however the long list of potential causes of LST errors emphasizes the difficulty of this
664 task. It is likely that most of these issues contribute in some way to LST errors and may even
665 compensate for each other.

666 Nevertheless we sum up the findings of the spatial LST validation of Mosaic, Noah and VIC and
667 attempt to identify potential causes to the spatial deficiencies of each LSM.

668 The case for Mosaic is the most unambiguous one. The general cool bias is more or less constant in
669 space and time and it can be attributed to an overestimation of aET. This finding is supported by the
670 positive annual evaporation bias at 961 small catchments over CONUS (Xia et al., 2012a) and a distinct
671 high bias in latent heat flux (Xia et al., 2012b) caused by vigorous upward water transport from the root
672 zone to the land surface (Mitchell et al., 2004). Generally, the connectivity of the cold phase is highest
673 for Mosaic among the three LSMs, which means that its cold patterns are smooth with clear transitions.
674 The overemphasized coupling between aET and LST might smoothen the simulated LST patterns,
675 because aET is controlled by the available energy, which naturally has smooth gradients. Future
676 research may not focus on improving Mosaic, because it can be regarded as a legacy model that will be
677 replaced in future NLDAS research.

678 Noah exhibits a quite similar LST pattern performance compared to Mosaic. However, in this case
679 errors in LST can clearly not be ascribed to aET errors (Figure 15). Xia et al. (2012b) identified
680 Noahs's higher albedo and its resulting lower net shortwave radiation as the reason for different LST
681 predictions in comparison to Mosaic. Noah has the highest mean monthly albedo for 10 out of 12
682 months among the three LSM over CONUS. The resulting lowest net shortwave radiation could
683 possibly explain Noah's cool bias. Recent works by Wei et al. (2013) and Xia et al. (2015b)
684 implemented improvements in the NLDAS-2 version of Noah. The emphasis was on adjusting the
685 roughness length for heat and the surface exchange coefficient to increase the aerodynamic
686 conductance, which yielded a significant improvement of the predicted LST patterns.

687 The spatial deficiencies in VIC are more complex than in the other LSMs. The lack of spatial
688 variability in the winter months is pointed out by the EOF analysis. This is due the distinct warm bias
689 in the northern part of CONUS (Figure 3) and can be attributed to an underestimation of aET. Further,
690 VIC's low connectivity for the cold phase stresses that the cold patterns are too heterogeneous, due to
691 the presence of disrupting warm cells. This spatial deficiency may be related to the occurrence of snow
692 but further analysis is needed to investigate this in more detail.

693 However, the reason why there is a general agreement between the LST patterns and their errors in
694 Mosaic and Noah, while VIC appears to have other controlling mechanisms of its LST patterns,
695 remains unanswered. To this end, further work is needed to better understand the drivers of spatial
696 variability of land surface variables with focus on their spatial patterns. There is clearly a demand for a
697 true spatial sensitivity analysis that can guide the modelling community on how to increase the spatial
698 pattern performance in LSMs.

5. Conclusion

This study provides a comprehensive spatial validation of three NLDAS-2 LSMs, namely Mosaic, Noah and VIC. A 30 year, satellite based (HIRS), LST dataset, suitable for monthly spatial validation of the annual cycle, is utilized to validate the models over CONUS. Although this study employs HIRS LST data only for CONUS the spatial coverage allows for global applications as well. Two innovative spatial performance metrics, namely an EOF analysis and a connectivity analysis, are applied to conduct a true pattern comparison, which goes beyond the standard cell to cell comparisons. We draw the following main conclusion from this work:

i. *Validation dataset:* The HIRS LST retrievals provide reasonable coverage over CONUS at a monthly aggregation level. The dataset has been validated against Fluxnet and BSRN stations and mostly warm biases are evident alongside a strong spatial and temporal correlation. The nature of the bias is complex in time and space and presumably caused, in part, by differences in spatial scales between the in-situ measurements and the satellite retrievals. This makes the HIRS LST dataset a suitable dataset for spatial LSM validations at large to global scales. However due to its uneven temporal distribution a validation is only meaningful at monthly time scale.

ii. *Spatial performance metrics:* The NLDAS-2 LSMs have distinct spatial and temporal biases and individual spatial model deficiencies that can be attributed to different causes. The joint EOF analysis of the observed and simulated LST maps is straightforward to interpret and combines the spatial and temporal component of the model validation. The first EOF captures more than 75% of the spatial variability and a strong spatial correlation is evident to the first EOF of air temperature. The second EOF adds an additional 6% of the explained variance and

addresses the seasonality of the LST patterns. Comparing the loadings for the observed and simulated LST maps allows us to derive a meaningful quantification of the spatial performance. For the first time, a connectivity analysis is applied to LST patterns and subsequently used as a spatial performance metric. It allows a separate analysis of the cold and warm patterns and shows that the LSMs are unable to simulate the complex pattern evolution during the transition from summer to winter. The LST patterns possess unique percolation thresholds, which strengthens the physical relevancy of connectivity as a characteristic of LST patterns. Connectivity, as a global measure with no local constrains, can be used to describe the homogeneity and smoothness of patterns. The RMSE between observed and simulated connectivity curves can quantify the spatial model performance. The inter-comparison of the spatial performance metrics by means of a correlation analysis underlines the difficulty of choosing a single, comprehensive metric. The metrics show redundant information depending on the nature of the spatial error. The connectivity of the cold LST patterns is the only metric that shows no redundancy to any other metrics and thus it clearly adds additional information to the validation that would be undetected by the other metrics.

iii. *Land atmosphere coupling:* Analyzing the complex coupling between daytime LST and aET helps to distinguish between water limited and energy limited conditions. Mosaic clearly performs worst at reproducing the observed coupling between the two variables while Noah is able to reproduce the coupling most accurately among the three LSMs. Overall, errors in LST are mostly related to errors in aET for Mosaic and VIC, but not for Noah. This emphasizes the usability of LST as a proxy to validate water balance errors in Mosaic and VIC.

Acknowledgments

The work has been carried out under the HOBE (Center for Hydrology in Denmark) and the SPACE (SPAtial Calibration and Evaluation in distributed hydrological modeling using satellite remote sensing data) project; both funded by the Villum foundation. We would like to acknowledge the NLDAS project for providing the simulated LST data. All LSM model output is freely available from the NLDAS homepage: http://ldas.gsfc.nasa.gov/nldas/. We would like to thank Lei Shi at NOAA National Climatic Data Center for providing the reprocessed HIRS satellite data. Also we would like to thank Fluxnet and AmeriFlux for providing high quality scientific flux data.

References

Alfieri, J. G., W. P. Kustas, J. H. Prueger, L. E. Hipps, J. L. Chávez, A. N. French and S. R. Evett, Intercomparison of nine micrometeorological stations during the BEAREX08 field campaign, Journal of Atmospheric and Oceanic Technology, 28(11), 1390-1406, 2011.

Anderson, M. C., W. P. Kustas, J. M. Norman, C. R. Hain, J. R. Mecikalski, L. Schultz, M. P. González-Dugo, C. Cammalleri, G. d'Urso and A. Pimstein, Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery, Hydrology and Earth System Sciences, 15(1), 223-239, 2011.

Andreadis, K. M., E. A. Clark, A. W. Wood, A. F. Hamlet and D. P. Lettenmaier, Twentieth-century drought in the conterminous United States, Journal of Hydrometeorology, 6(6), 985-1001, 2005.

Baldocchi, D., E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis and R. Evans, FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities, Bulletin of the American Meteorological Society, 82(11), 2415-2434, 2001.

Clark, M. P., Y. Fan, D. M. Lawrence, J. C. Adam, D. Bolster, D. J. Gochis, R. P. Hooper, M. Kumar, L. R. Leung and D. S. Mackay, Improving the representation of hydrologic processes in Earth System Models, Water Resources Research, 2015.

Cleugh, H. A., R. Leuning, Q. Mu and S. W. Running, Regional evaporation estimates from flux tower and MODIS satellite data, Remote sensing of Environment, 106(3), 285-304, 2007.

770 Coccia, G., A. L. Siemann, M. Pan and E. F. Wood, Creating consistent datasets by combining
771 remotely-sensed data and land surface model estimates through Bayesian uncertainty post-processing:
772 The case of Land Surface Temperature from HIRS, Remote sensing of Environment, 170, 290-305,
773 2015.

774 Corbari, C. and M. Mancini, Calibration and Validation of a Distributed Energy-Water Balance Model
775 Using Satellite Data of Land Surface Temperature and Ground Discharge Measurements, Journal of
776 Hydrometeorology, 15(1), 376-392, 2014.

777 dell Arciprete, D., R. Bersezio, F. Felletti, M. Giudici, A. Comunian and P. Renard, Comparison of
778 three geostatistical methods for hydrofacies simulation: a test on alluvial sediments, Hydrogeology
779 journal, 20(2), 299-311, 2012.

780 Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno and J. D. Tarpley,
781 Implementation of Noah land surface model advances in the National Centers for Environmental
782 Prediction operational mesoscale Eta model, Journal of Geophysical Research: Atmospheres (1984-
783 2012), 108(D22), 2003.

784 Famiglietti, J. S., D. Ryu, A. A. Berg, M. Rodell and T. J. Jackson, Field observations of soil moisture
785 variability across scales, Water Resources Research, 44(1), 2008.

786 Fang, Z., H. Bogena, S. Kollet, J. Koch and H. Vereecken, Spatio-temporal validation of long-term 3D
787 hydrological simulations of a forested catchment using empirical orthogonal functions and wavelet
788 coherence analysis, Journal of Hydrology, 2015.

789 Frey, C. M., C. Kuenzer, and S. Dech (2012), Quantitative comparison of the operational NOAA-
790 AVHRR LST product of DLR and the MODIS LST product V005, International journal of remote
791 sensing, 33(22), 7165-7183.

792 Getirana, A. C., E. Dutra, M. Guimberteau, J. Kam, H. Y. Li, B. Decharme, Z. Zhang, A. Ducharne, A.
793 Boone and G. Balsamo, Water balance in the Amazon basin from a land surface model ensemble,
794 Journal of Hydrometeorology, 15(6), 2586-2614, 2014.

795 Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati and E. E. Ebert, Intercomparison of spatial
796 forecast verification methods, Weather and Forecasting, 24(5), 1416-1430, 2009.

797 Graf, A., H. R. Bogena, C. Drüe, H. Hardelauf, T. Pütz, G. &. Heinemann and H. Vereecken,
798 Spatiotemporal relations between water budget components and soil water content in a forested
799 tributary catchment, Water Resources Research, 50(6), 4837-4857, 2014.

800 Grayson, R. B., G. Blőschl, A. W. Western and T. A. McMahon, Advances in the use of observed
801 spatial patterns of catchment hydrological response, Advances in Water Resources, 25(8), 1313-1334,
802 2002.

803 Gunshor, M. M., T. J. Schmit, and W. P. Menzel (2004), Intercalibration of the infrared window and
804 water vapor channels on operational geostationary environmental satellites using a single polar-orbiting
805 satellite, Journal of Atmospheric and Oceanic Technology, 21(1), 61-68.

806 Haddeland, I., Clark, D.B., Franssen, W., Ludwig, F., Voß, F., Arnell, N.W., Bertrand, N., Best, M.,
807 Folwell, S., Gerten, D. and Gomes, S., 2011. Multimodel estimate of the global terrestrial water
808 balance: Setup and first results. Journal of Hydrometeorology, 12(5), pp.869-884.

809 Heidinger, A. K., I. Laszlo, C. C. Molling, and D. Tarpley (2013), Using SURFRAD to verify the
810 NOAA single-channel land surface temperature algorithm, Journal of Atmospheric and Oceanic
811 Technology, 30(12), 2868-2884.

812 Hovadik, J. M. and D. K. Larue, Static characterizations of reservoirs: refining the concepts of
813 connectivity and continuity, Petroleum Geoscience, 13(3), 195-211, 2007.

814 Jackson, D. L., and B. J. Soden (2007), Detection and correction of diurnal sampling bias in HIRS/2
815 brightness temperatures, Journal of Atmospheric and Oceanic Technology, 24(8), 1425-1438.

816 Jackson, D. L., D. Wylie, and J. Bates (2003), The HIRS pathfinder radiance data set (1979–2001),
817 paper presented at Preprints, 12th Conf. on Satellite Meteorology and Oceanography, Long Beach, CA,
818 Amer. Meteor. Soc. P.

819 Jawson, S. D. and J. D. Niemann, Spatial patterns from EOF analysis of soil moisture at a large scale
820 and their dependence on soil, land-use, and topographic properties, Advances in Water Resources,
821 30(3), 366-381, 2007.

822 Jung, M., M. Reichstein, P. Ciais, S. I. Seneviratne, J. Sheffield, M. L. Goulden, G. Bonan, A. Cescatti,
823 J. Chen and R. De Jeu, Recent decline in the global land evapotranspiration trend due to limited
824 moisture supply, Nature, 467(7318), 951-954, 2010.

825 Karnieli, A., N. Agam, R. T. Pinker, M. Anderson, M. L. Imhoff, G. G. Gutman, N. Panov and A.
826 Goldberg, Use of NDVI and land surface temperature for drought assessment: merits and limitations,
827 Journal of Climate, 23(3), 618-633, 2010.

828 Koch, J., Z. Fang, T. Cornelissen, H. R. Bogena, B. Diekkrüger, S. Kollet and S. Stisen, Inter-
829 comparison of three distributed hydrological models with respect to seasonal variability of soil
830 moisture patterns at a small forested catchment., submitted to Journal of Hydrology, 2015a.

831 Koch, J., X. He, K. Jensen and J. C. Refsgaard, Challenges in conditioning a stochastic geological
832 model of a heterogeneous glacial aquifer to a comprehensive soft data set, Hydrology and Earth System
833 Sciences, 18(8), 2907-2923, 2014.

834 Koch, J., K. Jensen and S. Stisen, Toward a true spatial model evaluation in distributed hydrological
835 modeling: Kappa statistics, Fuzzy theory, and EOF-analysis benchmarked by the human perception and
836 evaluated against a modeling case study, Water Resources Research, 51(2), 1225-1246, 2015b.

837 Koirala, S., P. J. Yeh, Y. Hirabayashi, S. Kanae and T. Oki, Global-scale land surface hydrologic
838 modeling with the representation of water table dynamics, Journal of Geophysical Research:
839 Atmospheres, 119(1), 75-89, 2014.

840 Korres, W., C. N. Koyama, P. Fiener and K. Schneider, Analysis of surface soil moisture patterns in
841 agricultural landscapes using Empirical Orthogonal Functions, Hydrology and Earth System Sciences,
842 14(5), 751-764, 2010.

843 Korres, W., T. G. Reichenau, P. Fiener, C. N. Koyama, H. R. Bogena, T. Cornelissen, R. Baatz, M.
844 Herbst, B. Diekkrüger and H. Vereecken, Spatio-temporal soil moisture patterns-A meta-analysis
845 using plot to catchment scale data, Journal of Hydrology, 520, 326-341, 2015.

846 Koster, R. D. and M. J. Suarez, Modeling the land surface boundary in climate models as a composite
847 of independent vegetation stands, Journal of Geophysical Research: Atmospheres, 97(D3), 2697-2715,
848 1992.

849 Li, Z. l., B. H. Tang, H. Wu, H. Ren, G. Yan, Z. Wan, I. F. Trigo and J. A. Sobrino, Satellite-derived
850 land surface temperature: Current status and perspectives, Remote sensing of Environment, 131, 14-37,
851 2013.

852 Long, D., L. Longuevergne and B. R. Scanlon, Uncertainty in evapotranspiration from land surface
853 modeling, remote sensing, and GRACE satellites, Water Resources Research, 50(2), 1131-1151, 2014.

854 Mascaro, G., E. R. Vivoni and L. A. Méndez-Barroso, Hyperresolution hydrologic modeling in a
855 regional watershed and its interpretation using empirical orthogonal functions, Advances in Water
856 Resources, 83, 190-206, 2015.

857 McCabe, M. F. and E. F. Wood, Scale influences on the remote estimation of evapotranspiration using
858 multiple satellite sensors, Remote sensing of Environment, 105(4), 271-285, 2006.

859 Mitchell, K. E., D. Lohmann, P. R. Houser, E. F. Wood, J. C. Schaake, A. Robock, B. A. Cosgrove, J.
860 Sheffield, Q. Duan and L. Luo, The multi-institution North American Land Data Assimilation System
861 (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological
862 modeling system, Journal of Geophysical Research: Atmospheres, 109(D7), 2004.

863 Moradkhani, H., Hydrologic remote sensing and land surface data assimilation, Sensors, 8(5), 2986-
864 3004, 2008.

865 Mu, Q., M. Zhao and S. W. Running, Improvements to a MODIS global terrestrial evapotranspiration
866 algorithm, Remote sensing of Environment, 115(8), 1781-1800, 2011.

867 Ohmura, A., H. Gilgen, H. Hegner, G. Müller, M. Wild, E. G. Dutton, B. Forgan, C. Frőhlich, R.
868 Philipona and A. Heimo, Baseline Surface Radiation Network (BSRN/WCRP): New precision
869 radiometry for climate research, Bulletin of the American Meteorological Society, 79(10), 2115-2136,
870 1998.

871 Perry, M. A. and J. D. Niemann, Analysis and estimation of soil moisture at the catchment scale using
872 EOFs, Journal of Hydrology, 334(3), 388-404, 2007.

873 Qiu, J., X. Mo, S. Liu and Z. Lin, Exploring spatiotemporal patterns and physical controls of soil
874 moisture at various spatial scales, Theoretical and Applied Climatology, 118(1-2), 159-171, 2014.

875 Refsgaard, J. C., Parameterization, calibration and validation of distributed hydrological models,
876 Journal of Hydrology, 198(1-4), 69-97, 1997.

877 Refsgaard, J. C., Towards a formal approach to calibration and validation of models using spatial data,
878 Spatial patterns in catchment hydrology: observations and modelling, 329-354, 2001.

879 Reichle, R. H., S. V. Kumar, S. P. Mahanama, R. D. Koster and Q. Liu, Assimilation of satellite-
880 derived skin temperature observations into land surface models, Journal of Hydrometeorology, 11(5),
881 1103-1122, 2010.

882 Renard, P. and D. Allard, Connectivity metrics for subsurface flow and transport, Advances in Water
883 Resources, 51, 168-196, 2013.

884 Robel, J., NOAA KLM user's guide with NOAA-N,-P supplement, in Tech. rep., NOAA/National
885 Environmental Satellite, Data, and Information Services (NESDIS), 2009.

886 Rodell, M., Beaudoing, H.K., L'Ecuyer, T.S., Olson, W.S., Famiglietti, J.S., Houser, P.R., Adler, R.,
887 Bosilovich, M.G., Clayson, C.A., Chambers, D. and Clark, E., 2015. The observed state of the water
888 cycle in the early twenty-first century. Journal of Climate, 28(21), pp.8289-8318.

889

890 Sheffield, J., K. M. Andreadis, E. F. Wood and D. P. Lettenmaier, Global and continental drought in
891 the second half of the twentieth century: severity-area-duration analysis and temporal variability of
892 large-scale events, Journal of Climate, 22(8), 1962-1981, 2009.

893 Sheffield, J. and E. F. Wood, Characteristics of global and regional drought, 1950-2000: Analysis of
894 soil moisture data from off-line simulation of the terrestrial hydrologic cycle, Journal of Geophysical
895 Research: Atmospheres (1984-2012), 112(D17), 2007.

896 Sheffield, J., E. F. Wood, N. Chaney, K. Guan, S. Sadri, X. Yuan, L. Olang, A. Amani, A. Ali and S.
897 Demuth, A drought monitoring and forecasting system for sub-Sahara African water resources and
898 food security, Bulletin of the American Meteorological Society, 95(6), 861-882, 2014.

899 Shi, L., Global atmospheric temperature and humidity profiles based on intersatellite calibrated HIRS
900 measurements, 2011.

901 Shi, L. and J. J. Bates, Three decades of intersatellite-calibrated High-Resolution Infrared Radiation
902 Sounder upper tropospheric water vapor, Journal of Geophysical Research: Atmospheres (1984-2012),
903 116(D4), 2011.

Siemann, A., G. Coccia, M. Pang and E. F. Wood, Development and Analysis of a Long Term, High-Resolution, Global, Terrestrial Land Surface Temperature Dataset., Journal of Climate, 2016.

Silvestro, F., S. Gabellani, F. Delogu, R. Rudari and G. Boni, Exploiting remote sensing land surface temperature in distributed hydrological modelling: the example of the Continuum model, Hydrology and Earth System Sciences, 17(1), 39-62, 2013.

Stisen, S., M. F. McCabe, J. C. Refsgaard, S. Lerer and M. B. Butts, Model parameter analysis using remotely sensed pattern information in a multi-constraint framework, Journal of Hydrology, 409(1), 337-349, 2011.

Stisen, S., I. Sandholt, A. Nørgaard, R. Fensholt and K. H. g. Jensen, Combining the triangle method with thermal inertia to estimate regional evapotranspiration-Applied to MSG-SEVIRI data in the Senegal River basin, Remote sensing of Environment, 112(3), 1242-1255, 2008.

Stoy, P. C., M. Mauder, T. Foken, B. Marcolla, E. Boegh, A. Ibrom, M. A. Arain, A. Arneth, M. Aurela, and C. Bernhofer (2013), A data-driven analysis of energy balance closure across FLUXNET research sites: The role of landscape scale heterogeneity, Agricultural and Forest Meteorology, 171, 137-152.

Sun, D. and R. T. Pinker, Estimation of land surface temperature from a Geostationary Operational Environmental Satellite (GOES-8), Journal of Geophysical Research: Atmospheres (1984-2012), 108(D11), 2003.

Trigo, I. F., I. T. Monteiro, F. Olesen and E. Kabsch, An assessment of remotely sensed land surface temperature, Journal of Geophysical Research: Atmospheres (1984-2012), 113(D17), 2008.

924 Troy, T. J., J. Sheffield and E. F. Wood, Estimation of the terrestrial water budget over northern
925 Eurasia through the use of multiple data sources, Journal of Climate, 24(13), 3272-3293, 2011.

926 Velpuri, N. M., G. B. Senay, R. K. Singh, S. Bohms and J. P. Verdin, A comprehensive evaluation of
927 two MODIS evapotranspiration products over the conterminous United States: Using point and gridded
928 FLUXNET and water balance ET, Remote sensing of Environment, 139, 35-49, 2013.

929 Wan, Z., Y. Zhang, Q. Zhang and Z. l. Li, Validation of the land-surface temperature products
930 retrieved from Terra Moderate Resolution Imaging Spectroradiometer data, Remote sensing of
931 Environment, 83(1), 163-180, 2002.

932 Wan, Z., Y. Zhang, Q. Zhang, and Z.-L. Li (2004), Quality assessment and validation of the MODIS
933 global land surface temperature, International Journal of Remote Sensing, 25(1), 261-274.

934 Wanders, N., M. F. Bierkens, S. M. de Jong, A. de Roo and D. Karssenberg, The benefits of using
935 remotely sensed soil moisture in parameter identification of large-scale hydrological models, Water
936 Resources Research, 50(8), 6874-6891, 2014.

937 Wang, K. and S. Liang, Evaluation of ASTER and MODIS land surface temperature and emissivity
938 products using long-term surface longwave radiation observations at SURFRAD sites, Remote sensing
939 of Environment, 113(7), 1556-1565, 2009.

940 Wang, L., T. Koike, K. Yang and P. J.-F. Yeh, Assessment of a distributed biosphere hydrological
941 model against streamflow and MODIS land surface temperature in the upper Tone River Basin, Journal
942 of Hydrology, 377(1), 21-34, 2009.

943 Wealands, S. R., R. B. Grayson and J. P. Walker, Quantitative comparison of spatial fields for
944 hydrological model assessment - Some promising approaches, 28(1), 15-32, 2005.

945 Wei, H., Y. Xia, K. E. Mitchell and M. B. Ek, Improvement of the Noah land surface model for warm
946 season processes: Evaluation of water and energy flux simulation, Hydrological Processes, 27(2), 297-
947 303, 2013.

948 Western, A. W., G. Bloeschl and R. B. Grayson, Toward capturing hydrologically significant
949 connectivity in spatial patterns, Water Resources Research, 37(1), 83-97, 2001.

950 Wilson, K., et al. (2002), Energy balance closure at FLUXNET sites, Agricultural and Forest
951 Meteorology, 113(1–4), 223-243.

952 Wolff, J. K., M. Harrold, T. Fowler, J. H. Gotway, L. Nance and B. G. Brown, Beyond the basics:
953 Evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods,
954 Weather and Forecasting, 29(6), 1451-1472, 2014.

955 Wood, E. F., D. Lettenmaier, X. Liang, B. Nijssen and S. W. Wetzel, Hydrological modeling of
956 continental-scale basins, Annual Review of Earth and Planetary Sciences, 25(1), 279-300, 1997.

957 Wood, E. F., J. K. Roundy, T. J. Troy, L. P. H. Van Beek, M. F. Bierkens, E. Blyth, A. de Roo, P. Dőll,
958 M. Ek and J. Famiglietti, Hyperresolution global land surface modeling: Meeting a grand challenge for
959 monitoring Earth's terrestrial water, Water Resources Research, 47(5), 2011.

960 Wylie, D., D. L. Jackson, W. P. Menzel, and J. J. Bates (2005), Trends in global cloud cover in two
961 decades of HIRS observations, Journal of climate, 18(15), 3021-3031.

962 Xia, Y., M. Ek, J. Sheffield, B. Livneh, M. Huang, H. Wei, S. Feng, L. Luo, J. Meng and E. Wood,
963 Validation of Noah-simulated soil temperature in the North American land data assimilation system
964 phase 2, Journal of Applied Meteorology and Climatology, 52(2), 455-471, 2013.

965 Xia, Y., M. T. Hobbins, Q. Mu and M. B. Ek, Evaluation of NLDAS-2 evapotranspiration against
966 tower flux site observations, Hydrological Processes, 29(7), 1757-1771, 2015a.

967 Xia, Y., K. Mitchell, M. Ek, B. Cosgrove, J. Sheffield, L. Luo, C. Alonge, H. Wei, J. Meng and B.
968 Livneh, Continental-scale water and energy flux analysis and validation for North American Land Data
969 Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow,
970 Journal of Geophysical Research: Atmospheres, 117(D3), 2012a.

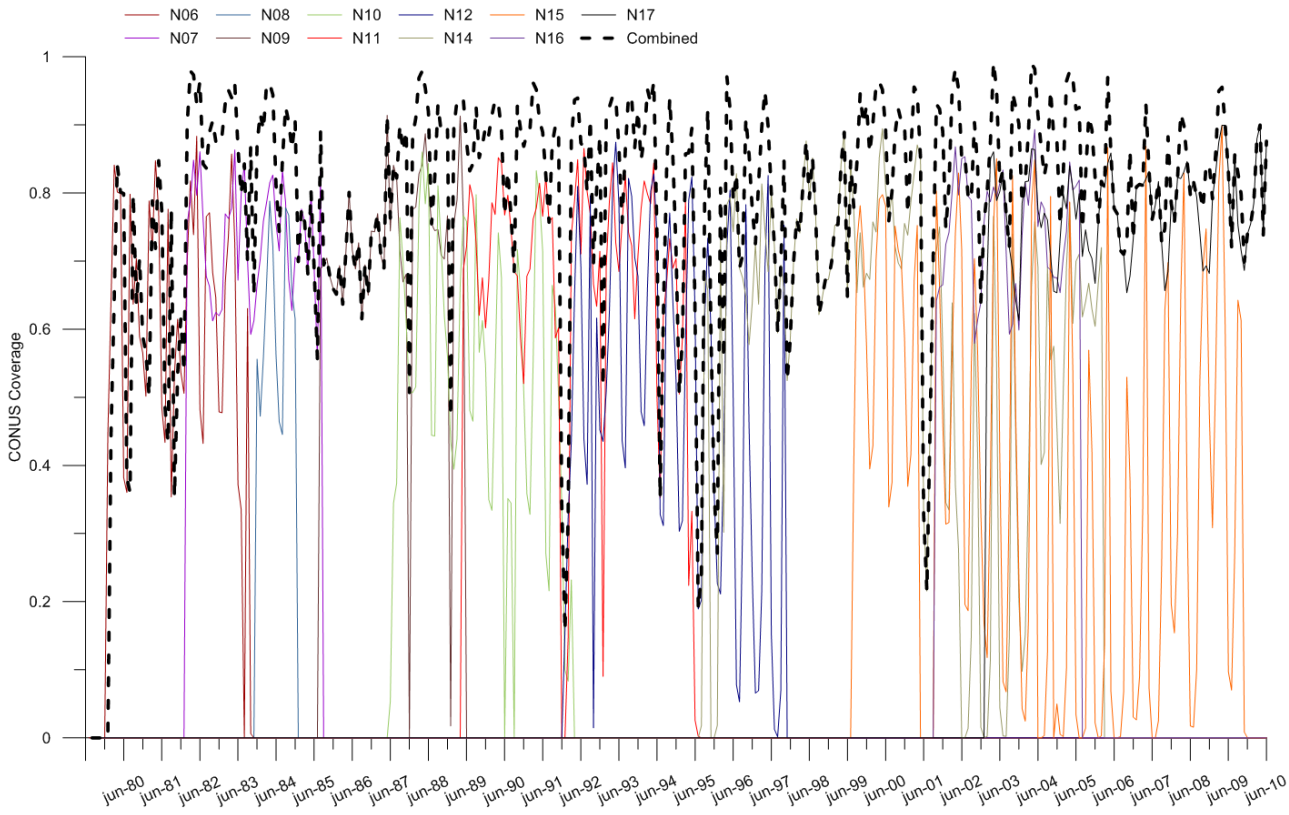
971 Xia, Y., K. Mitchell, M. Ek, J. Sheffield, B. Cosgrove, E. Wood, L. Luo, C. Alonge, H. Wei and J.
972 Meng, Continental-scale water and energy flux analysis and validation for the North American Land
973 Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model
974 products, Journal of Geophysical Research: Atmospheres, 117(D3), 2012b.

975 Xia, Y., C. D. Peter-Lidard, M. Huang, H. Wei and M. Ek, Improved NLDAS-2 Noah-simulated
976 hydrometeorological products with an interim run, Hydrological Processes, 29(5), 780-792, 2015b.

977 Xia, Y., J. Sheffield, M. B. Ek, J. Dong, N. Chaney, H. Wei, J. Meng and E. F. Wood, Evaluation of
978 multi-model simulated soil moisture in NLDAS-2, Journal of Hydrology, 512, 107-125, 2014.

979

980 **List of Figures**



981

982 Figure 1. Monthly coverage of HIRS LST retrievals over CONUS for each of the 11 NOAA satellites

983 and the combined coverage from July 1979 to July 2009.

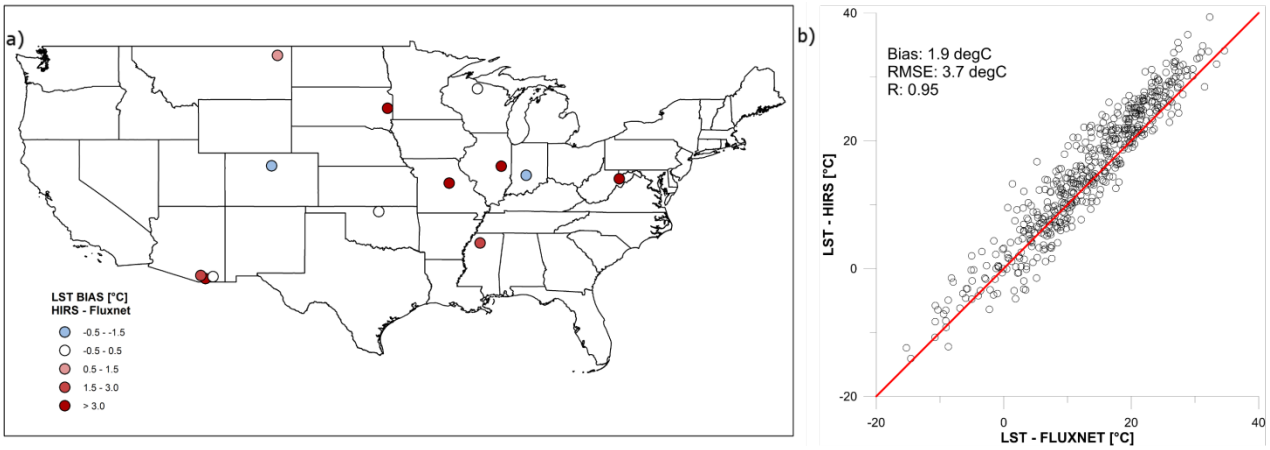
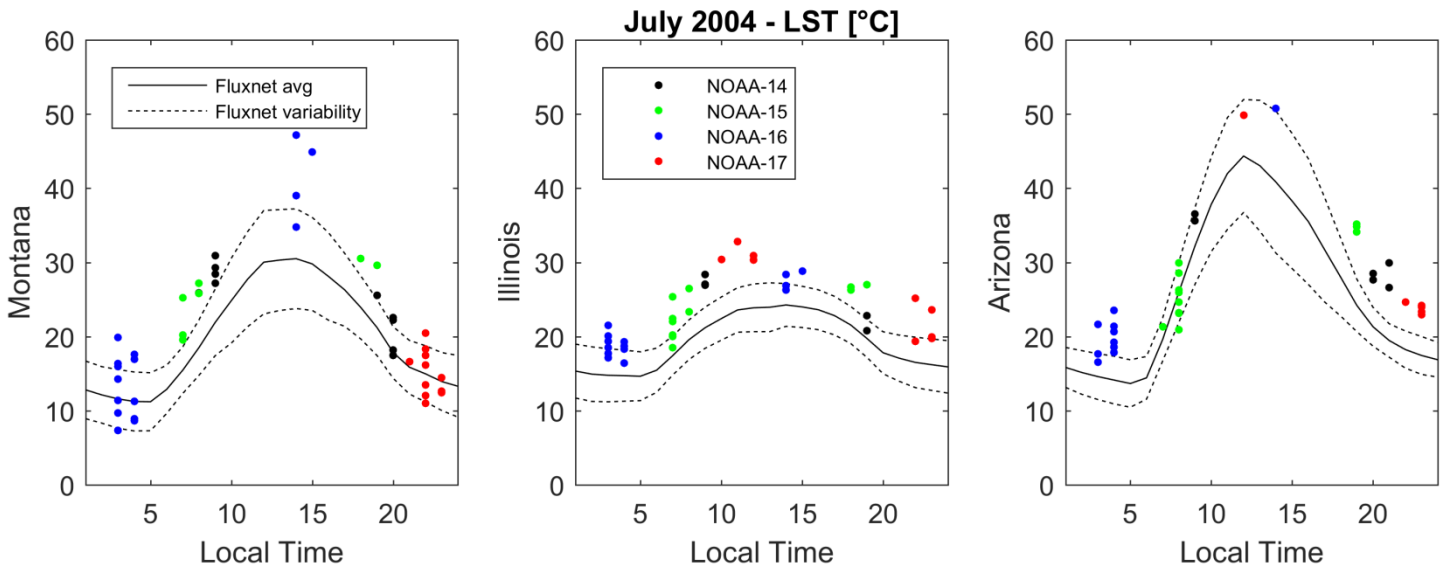


Figure 2. Comparison of monthly LST data between Fluxnet and HIRS at 15 stations over CONUS. a) The bias at each station that has at least one full year of data and b) Scatter plot for all stations (511 months at 15 stations).

989



990

991 Figure 3. Diurnal validation of HIRS LST against Fluxnet data at three sites across the US: Fort Peck in
992 Montana, Bondville in Illinois and Audubon Research Ranch in Arizona. Fluxnet observations are
993 averaged for each hour for July 2004 and the variability is expressed by +/- one standard deviation.
994 Each individual HIRS observations from July 2004 at the collocated 0.5 degree grid is included in the
995 figure.

996

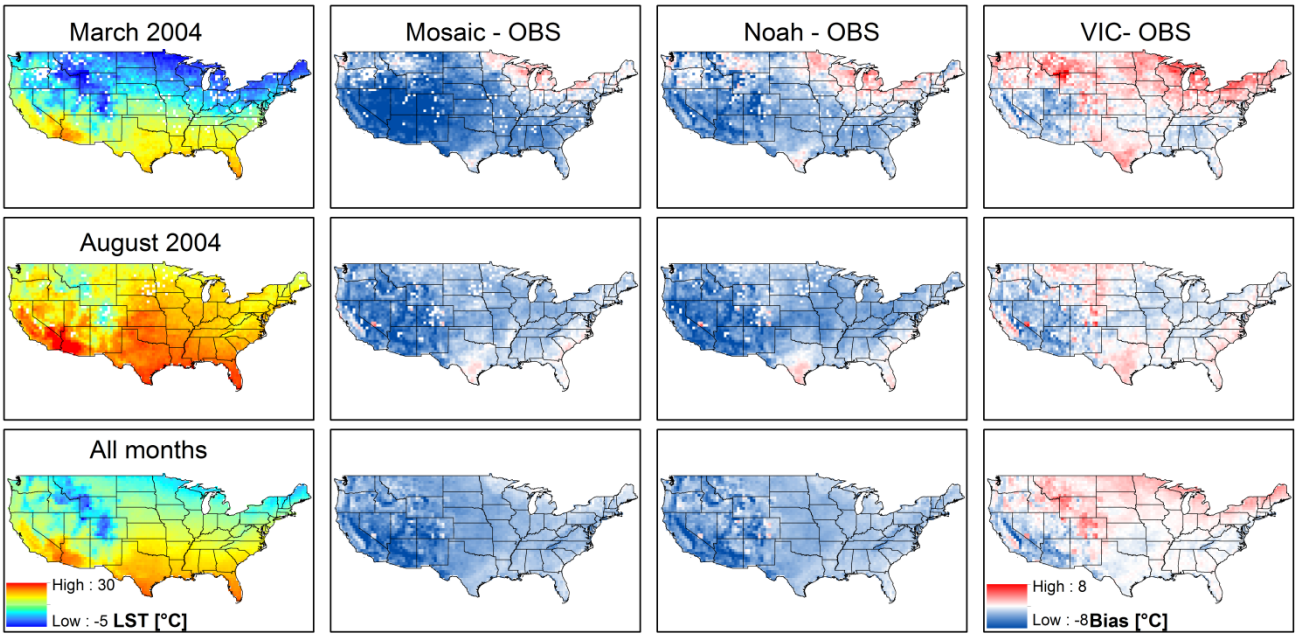
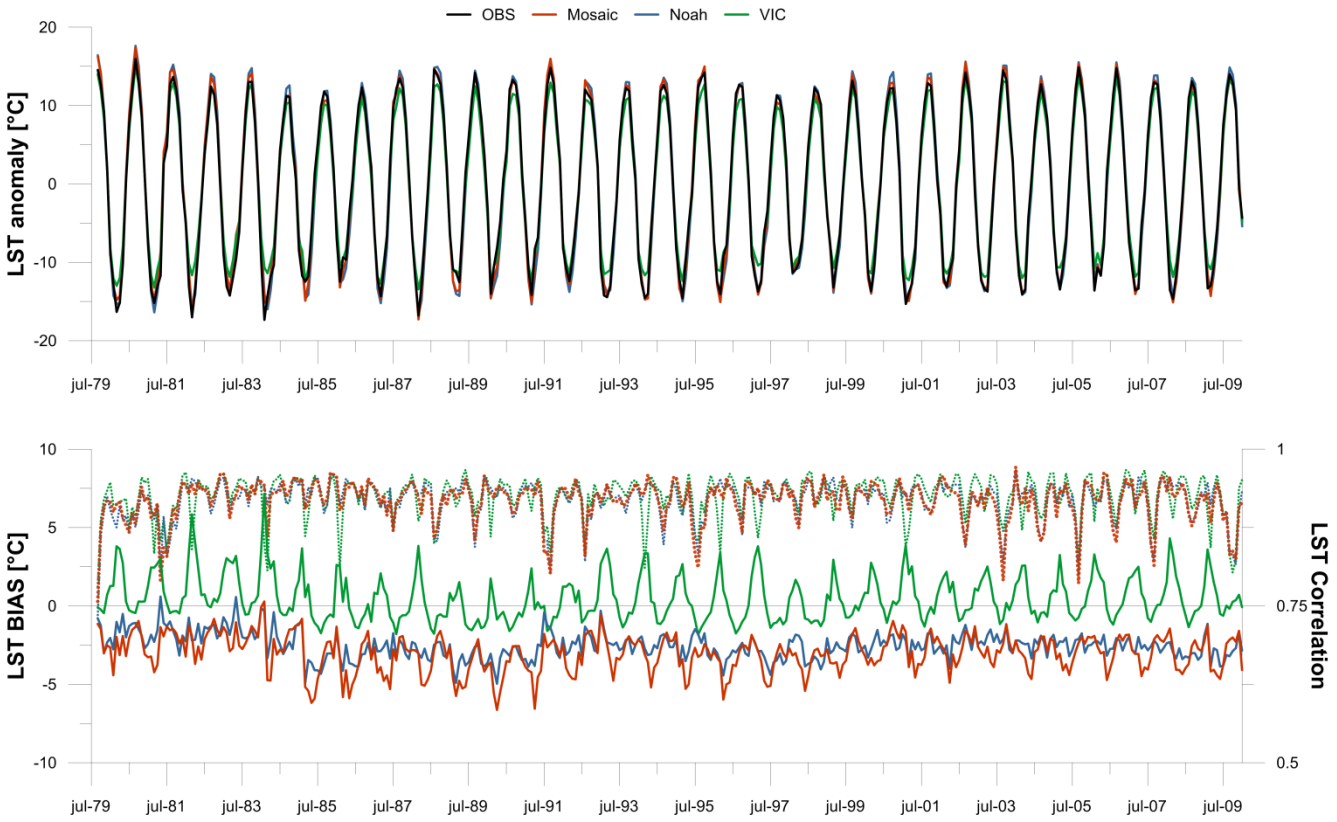


Figure 4. The left column presents observed (HIRS) LST maps for March 2004, August 2004 and the average of all months. Column two to four show the LST residuals for Mosaic, Noah and VIC, respectively. Red colors indicate a warm bias and cold colors indicate a cold bias.

1001

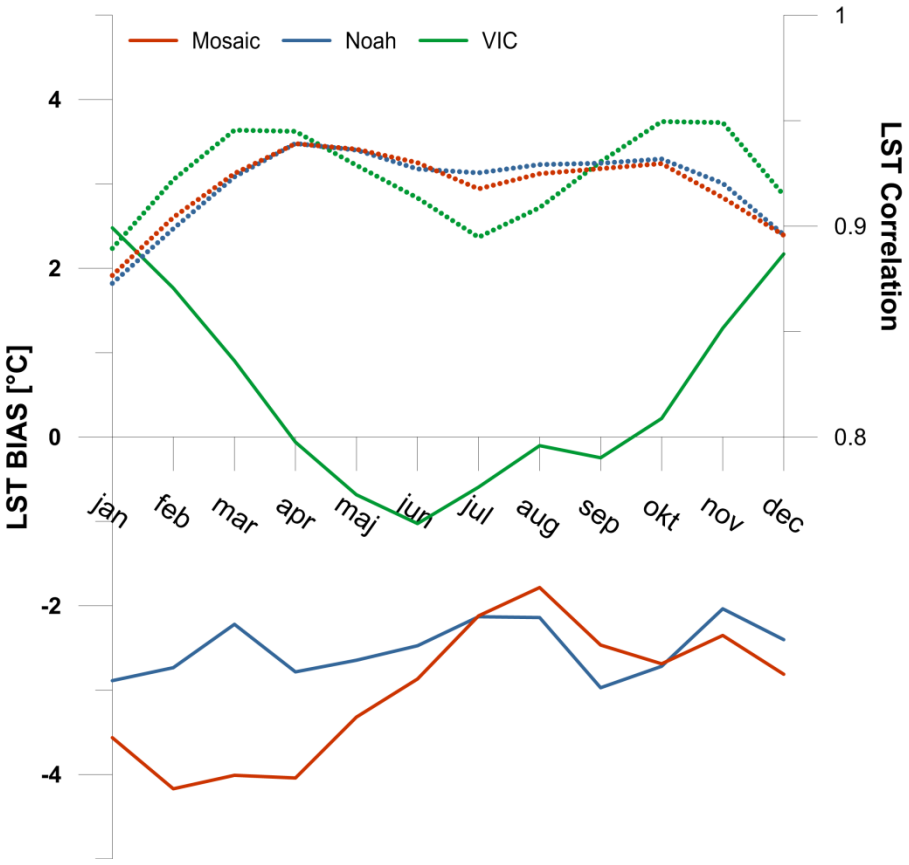


1002

1003 Figure 5. The top panel depicts the monthly variation of the observed (HIRS) and simulated (Mosaic,

1004 NOAH and VIC) monthly mean LST anomaly. The bottom panel presents the monthly LST bias (LSM

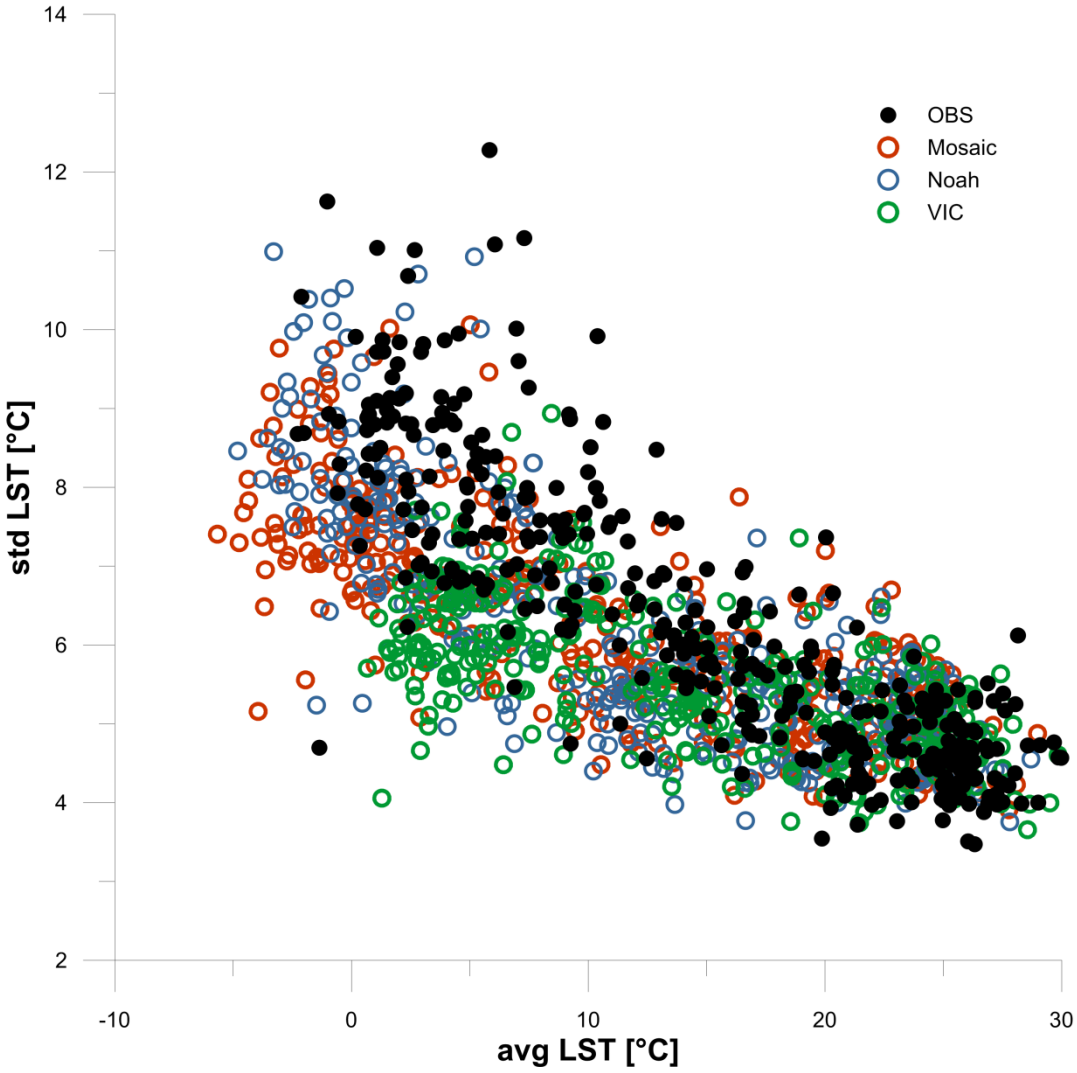
1005 - HIRS; solid line) and the monthly spatial correlation (dotted line) for the three LSMs.



1006

1007 Figure 6. The average monthly LST bias (LSM - HIRS; solid line) and the average monthly spatial
1008 correlation (dotted line) for the 30 year period.

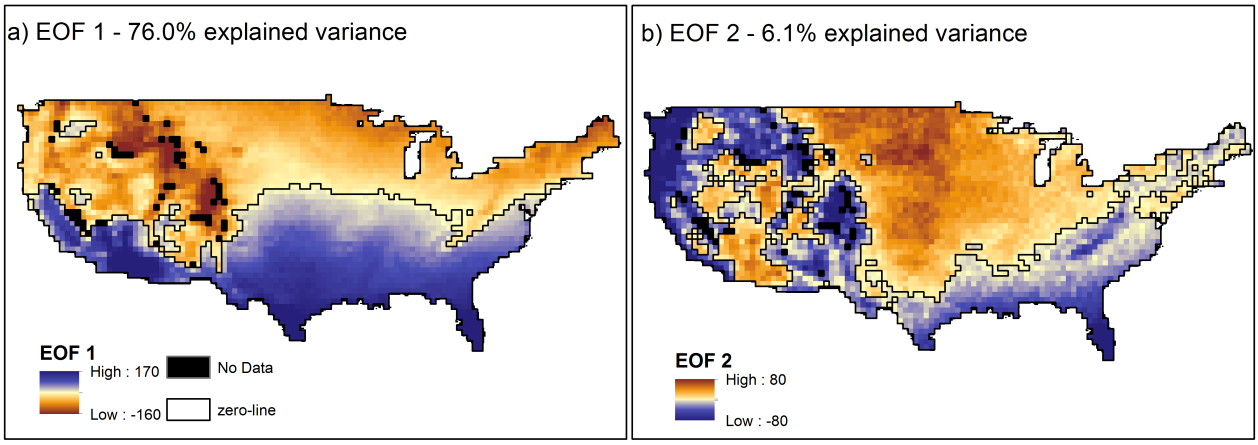
1009



1010

1011 Figure 7. The spatio-temporal variability of LST from HIRS and the three LSMs depicted by the mean
1012 monthly LST versus the monthly spatial variability of LST (standard deviation).

1013

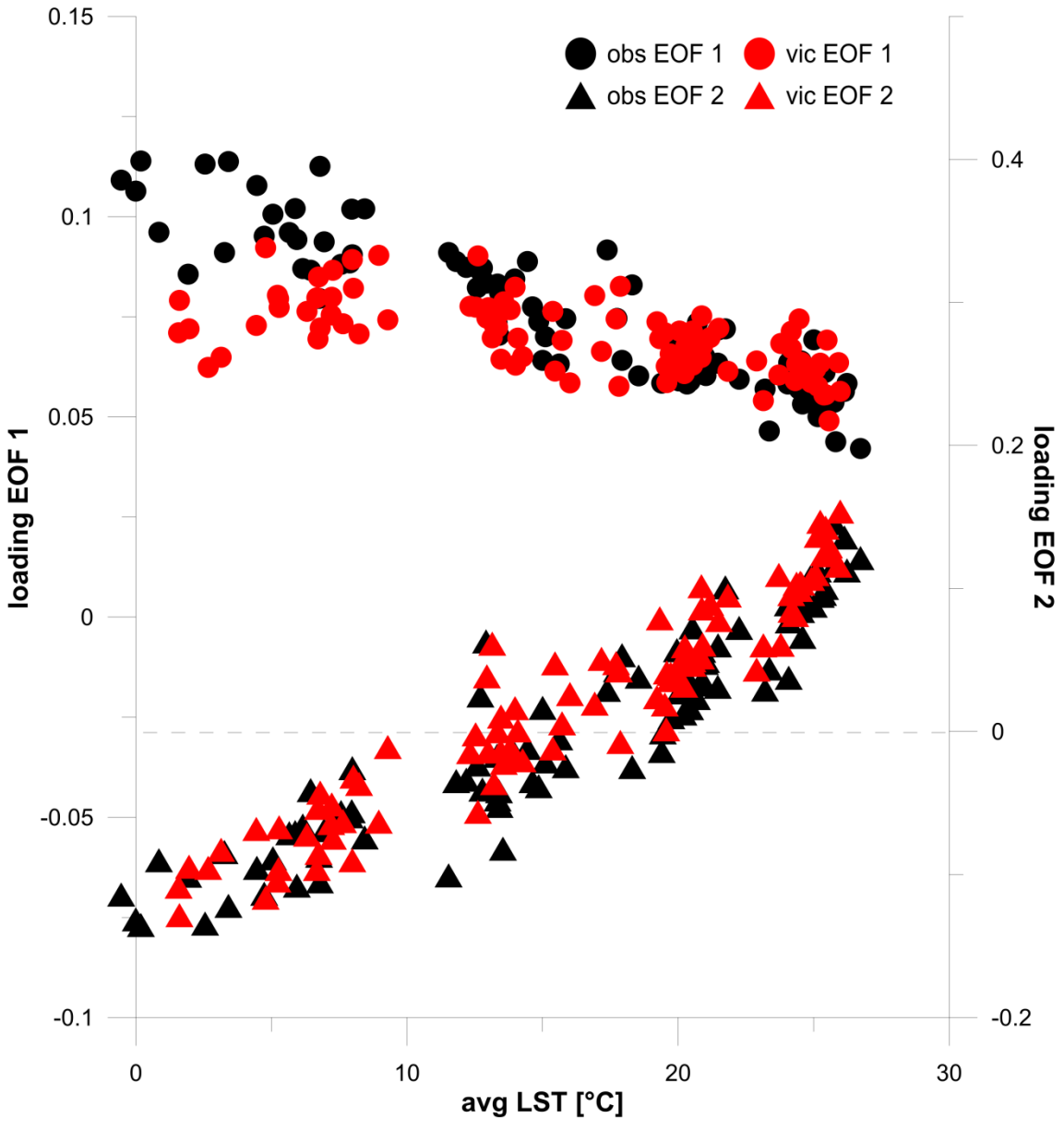


1014

1015 Figure 8. The resulting maps for EOF 1 and EOF 2 based on the joint EOF-analysis of 91 (coverage >

1016 0.9) monthly HIRS and VIC LST maps.

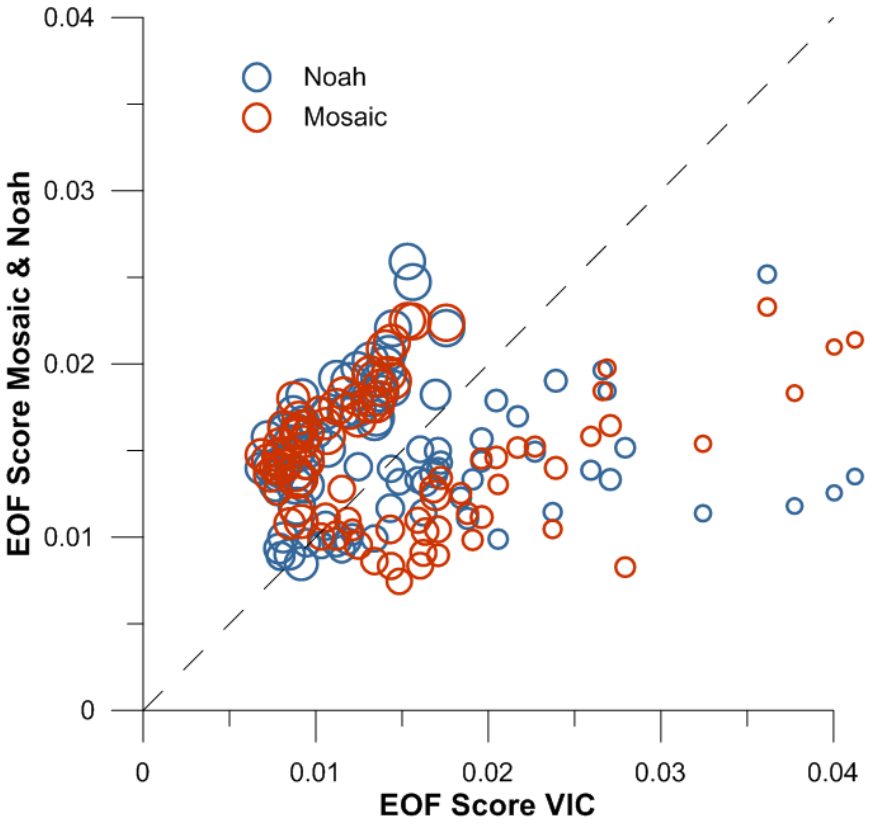
1017



1018

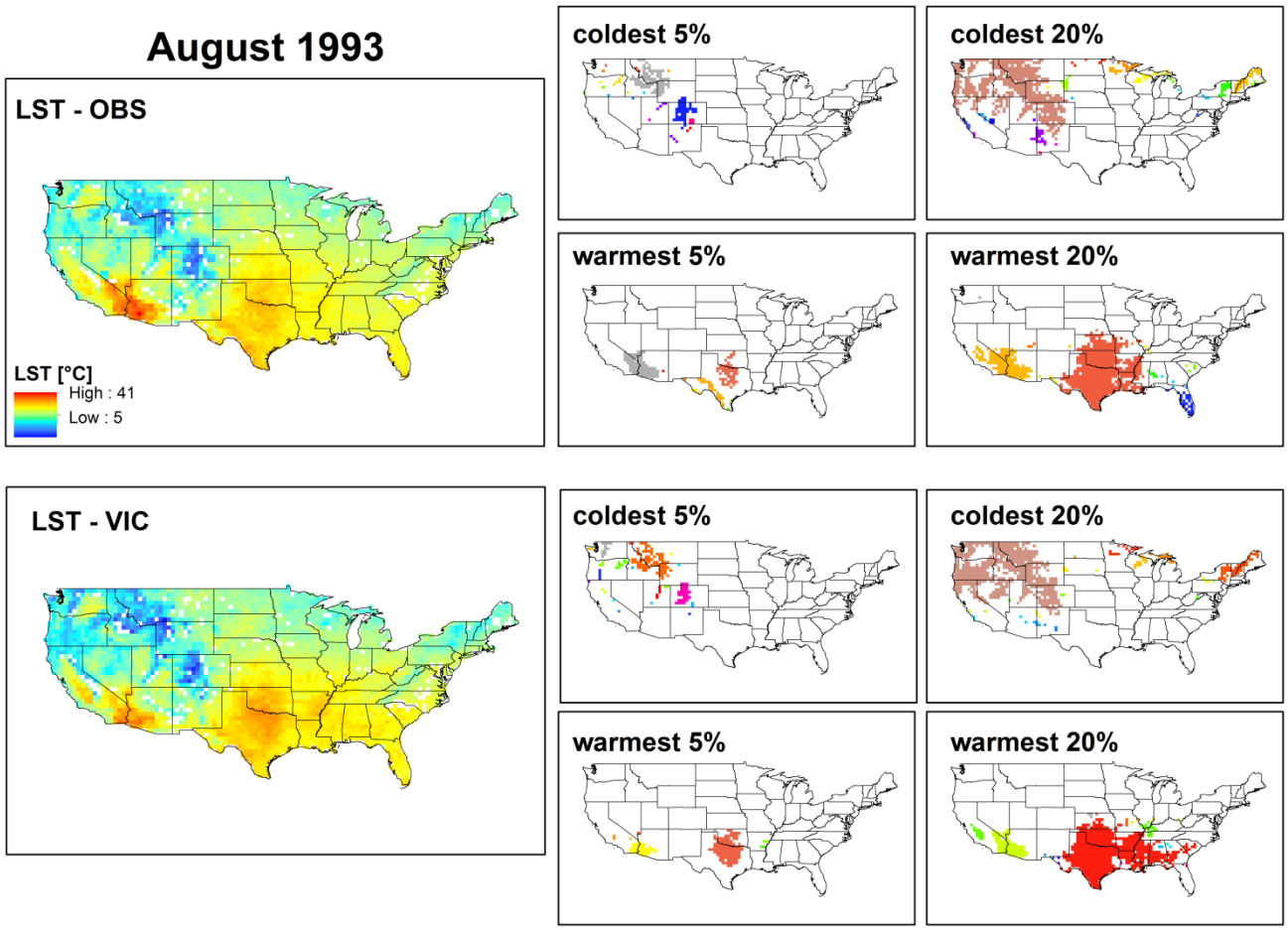
1019 Figure 9. The resulting loadings for EOF 1 and EOF 2 based on the joint EOF-analysis of 91 (coverage

1020 > 0.9) monthly HIRS (obs) and VIC LST maps plotted against the average monthly LST.



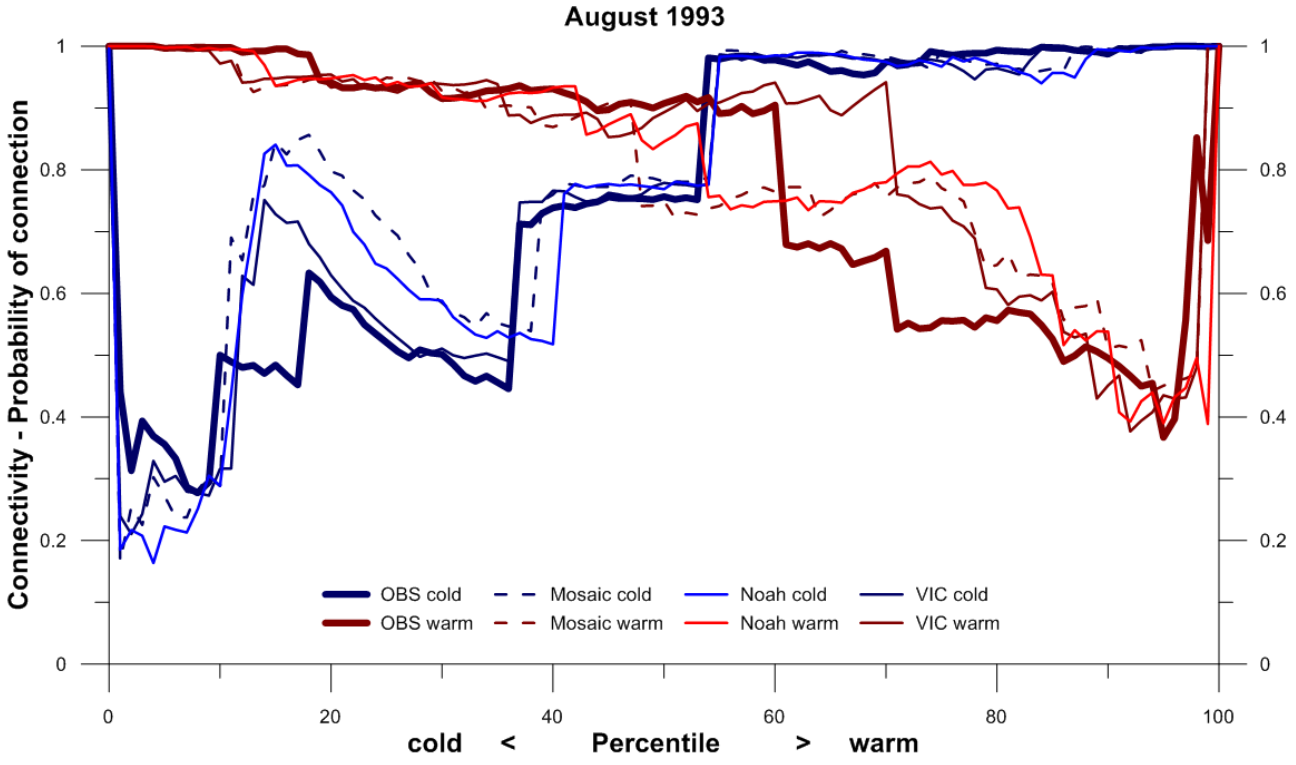
1021

1022 Figure 10. Scatterplot showing the comparison of the EOF based performance metric for the three
1023 LSMs for the 91 months with a coverage greater than 0.9. The lower the score the better the spatial
1024 performance. The size of the circles represents the average monthly CONUS LST given by HIRS;
1025 ranging from -0.5°C (smallest circle) to 26.7°C (largest circle).



1026

1027 Figure 11. An example of the connectivity analysis of observed (HIRS) and simulated (VIC) LST maps
1028 for August 1993. The left panel shows the original LST maps and the right panel presents the results
1029 from the cluster analysis for the coldest and warmest 5% and 20% of the cells. Each connected cluster
1030 is assigned a unique color.



1031

1032 Figure 12. The connectivity, quantified by the probability of connection, for the warm phase (red) and
1033 cold phase (blue) for August 1993. The probability of connection is computed at all percentiles that
1034 truncate the continuous LST maps into binary (cold/warm) maps.

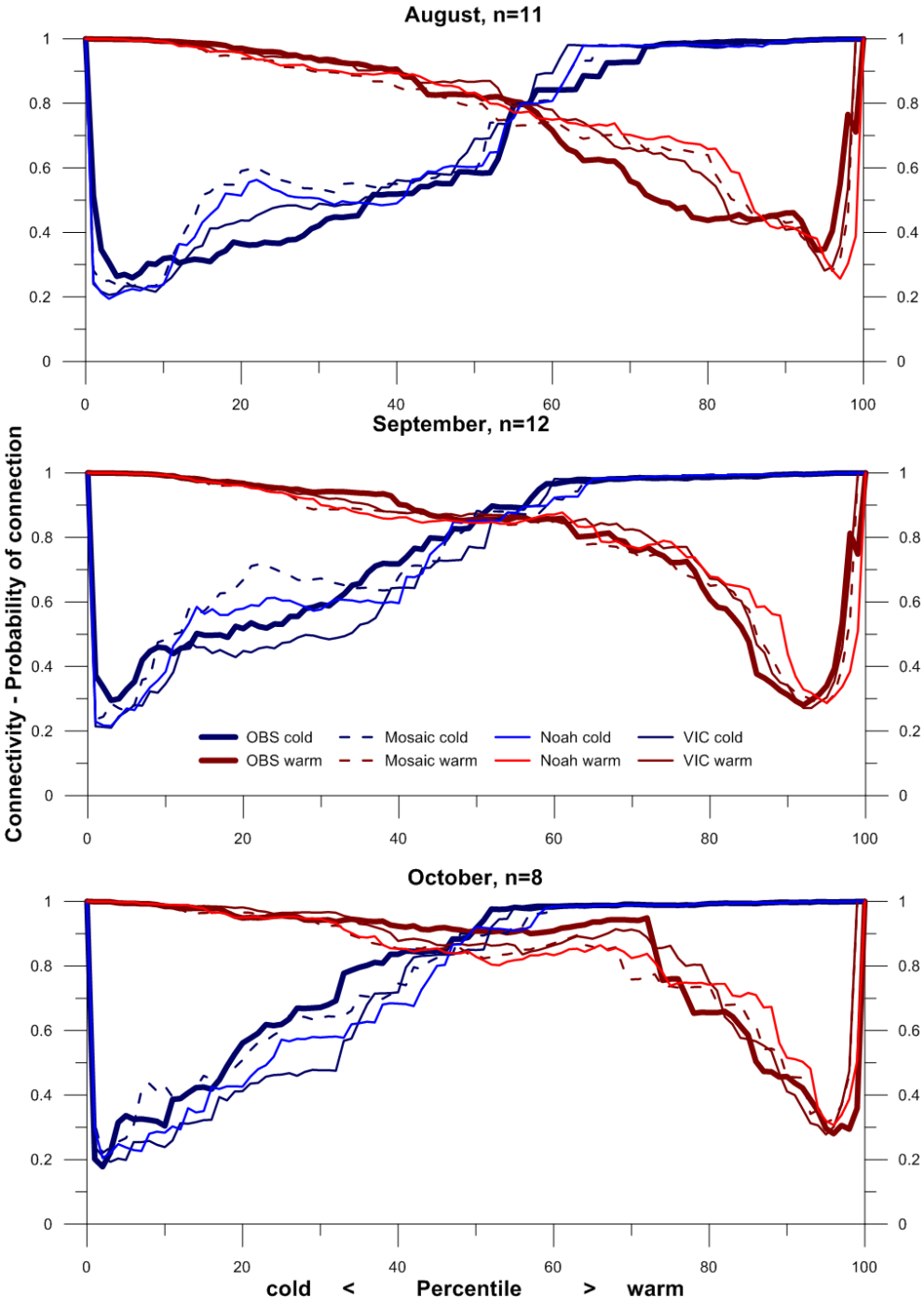


Figure 13. Average connectivity curves for August, September and October. The connectivity-analysis is conducted for 33 months where the coverage is greater than 0.95. These months are predominantly August (11), September (12) and October (8).

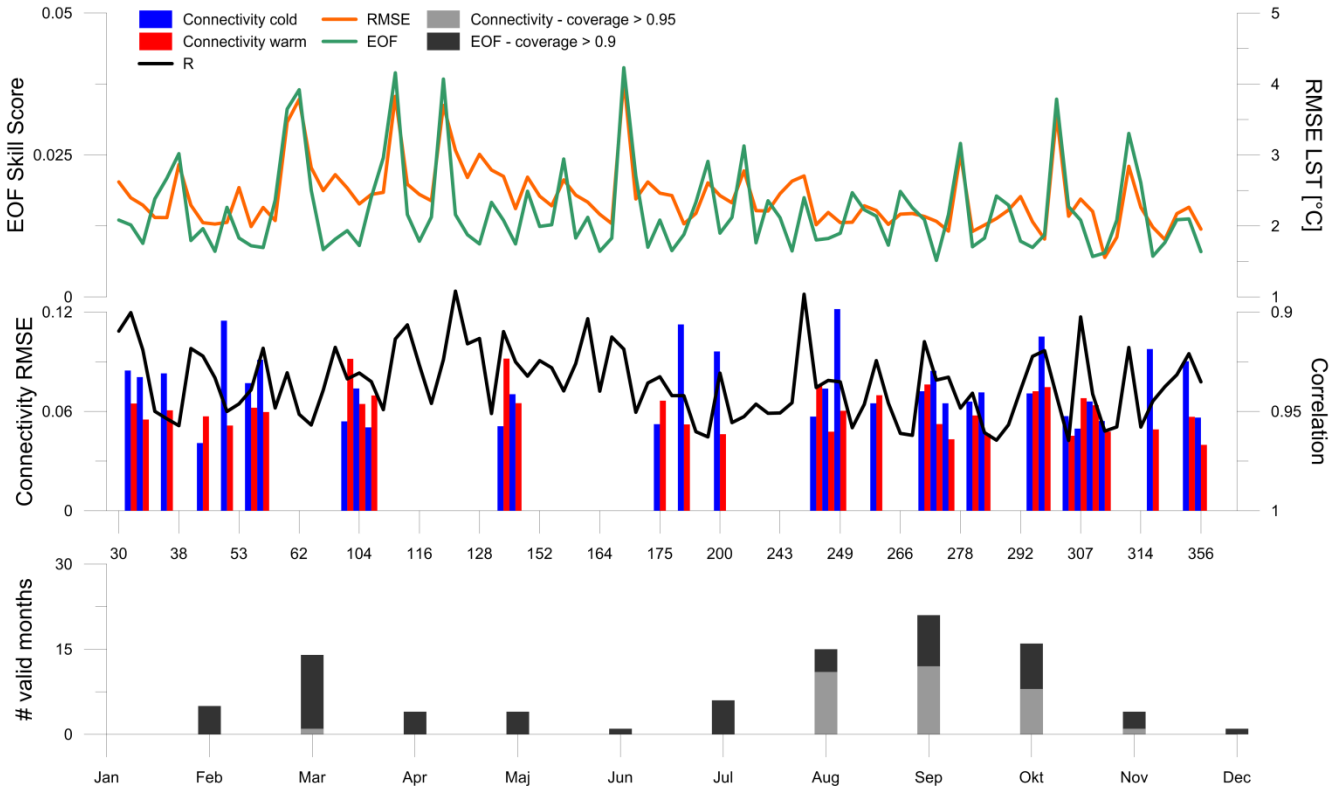
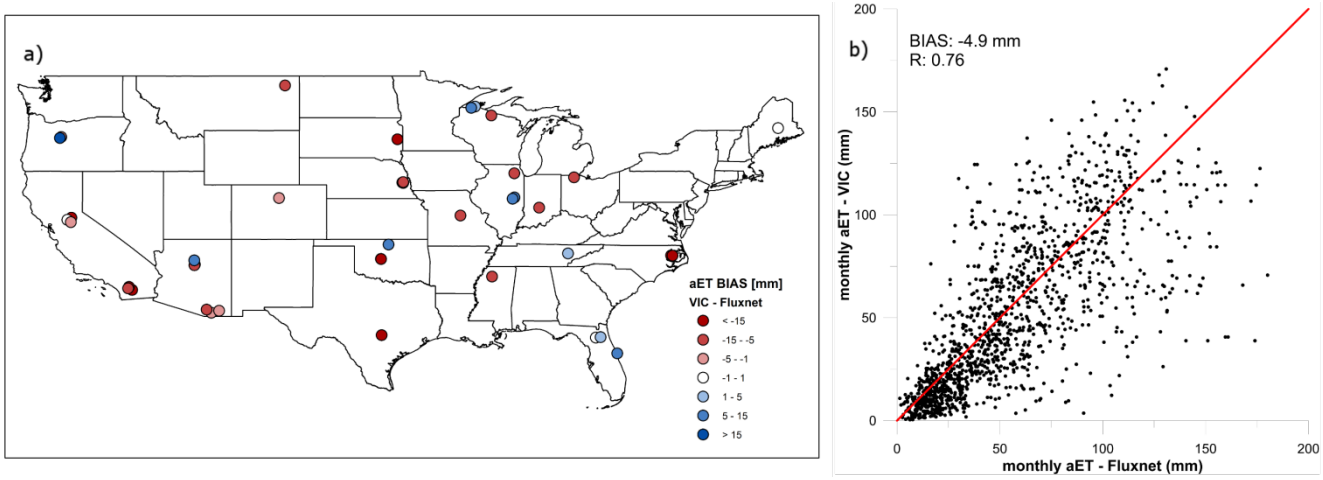


Figure 14. The top and middle panel shows a comparison of various spatial performance metrics: EOF analysis, connectivity analysis, root-mean-squared-error (RMSE) and spatial correlation (R). The results are only shown for the 91 months with a coverage greater than 0.9 that are used for the EOF-analysis, thus the X-axis is not equidistant in time. The connectivity analysis is only conducted for months with a coverage greater than 0.95 (33 months). The bottom panel illustartes the distribution of months used for the EOF analysis and the connectivity analysis.

1046



1047

1048 Figure 15. Comparison of monthly aET data between Fluxnet and VIC at 41 stations over CONUS. a)
1049 depicts the bias at each station that has at least on full year of data and b) combines all available data at
1050 all stations into one scatter plot (1311 months at 41 Fluxnet sites).

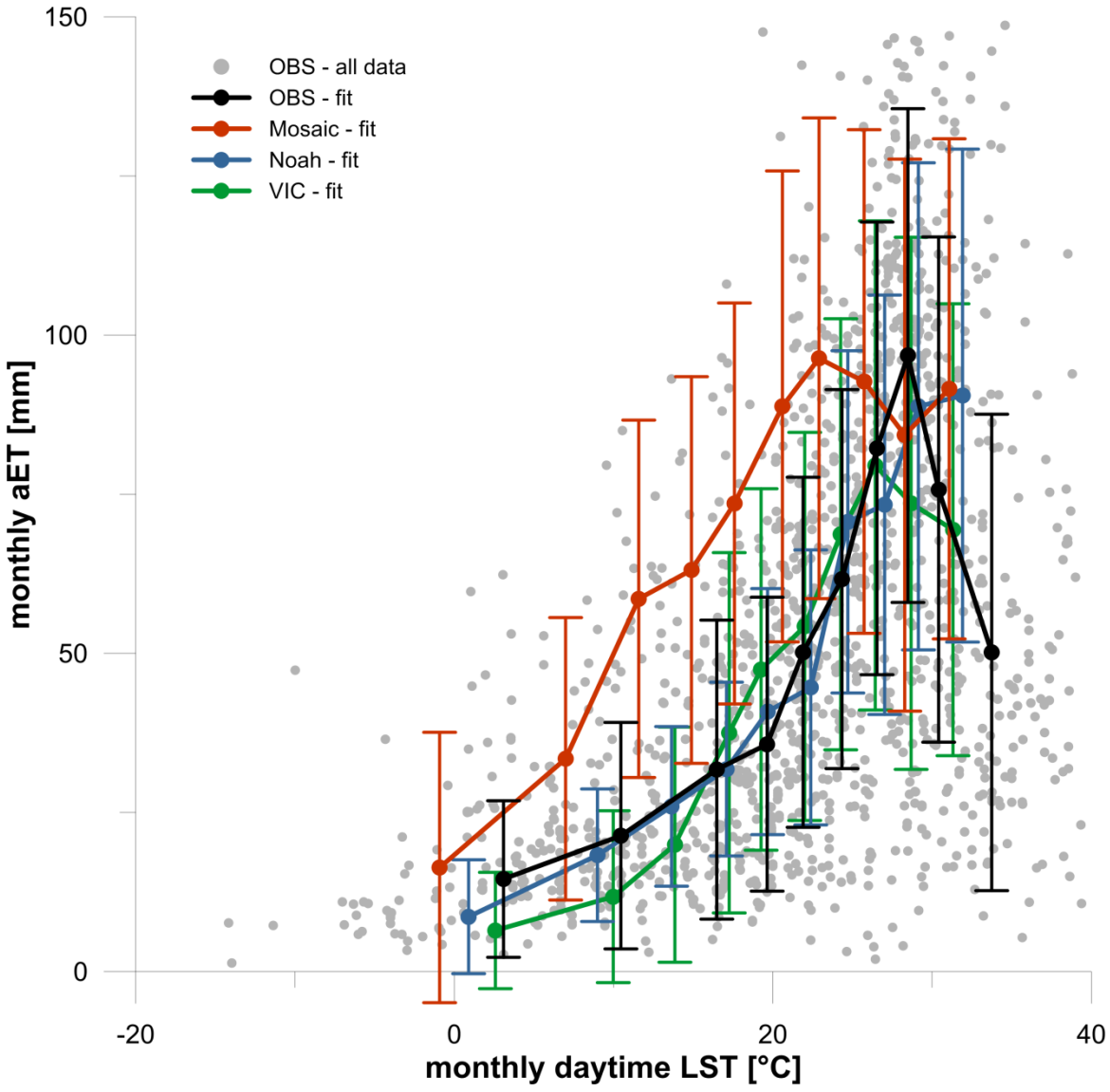


Figure 16. The coupling between monthly averages of daytime LST and monthly sums of aET for the observed data (aET from Fluxnet and LST from HIRS) and purely modelled data (1311 months at 41 Fluxnet sites). The fitted curves are based on grouping the data into 10 equally sized bins following the LST percentiles; the points represent mean LST and mean aET per bin and the error bar represents the standard deviation of aET per bin.

List of tables

Tabel 1. Comparison of various spatial performance metrics, EOF analysis, connectivity analysis, root-mean-squared-error (RMSE) and spatial correlation (R), on the basis of their correlation coefficients. A strong correlation between two metrics indicates that they provide redundant information. Strong correlations (> 0.5) are highlighted.

		RMSE	R	EOF	Con-cold	Con-warm
Mosaic	RMSE	1.0	0.0	-0.3	-0.1	-0.2
	R		1.0	**-0.6**	-0.4	**-0.5**
	EOF			1.0	0.4	**0.6**
	Con-cold				1.0	0.2
	Con-warm					1.0
Noah	RMSE	1.0	-0.1	0.1	-0.1	0.2
	R		1.0	**-0.6**	0.0	**-0.5**
	EOF			1.0	0.0	**0.6**
	Con-cold				1.0	0.2
	Con-warm					1.0
VIC	RMSE	1.0	-0.2	**0.8**	-0.2	**0.6**
	R		1.0	0.1	0.1	**-0.5**
	EOF			1.0	0.0	0.3
	Con-cold				1.0	-0.3
	Con-warm					1.0