Dear Authors,

Thank you for uploading the revised version of this manuscript. There are still some remaining minor comments.

Best,
Shrad

We are thankful to the Editor for these comments. We have addressed the recommended minor changes as detailed below.

Section 3.1: Please change "employ the WFDEI dataset (Watch Forcing Data methodology applied to Era-Interim data, Weedon et al. 2014)" to Watch Forcing Data methodology applied to Era-Interim (WFDEI, Weedon et al. 2014) dataset.

A1: Done.

In response to Reviewer #2, comment #7, I do not see a map showing soil moisture and streamflow stations in the manuscript. Likewise, I also do not see the Table S1. Perhaps you have uploaded an older version of the supplementary material?

A2: We will make sure to upload the updated supplementary material which includes the map and Table S1.

In response to Reviewer #3, comment #10 on section 3.1. What additional information regarding the forcing datasets has the authors included that was not present in the previous version. As far as I can tell the only difference in this version with the previous version is that this version contains the full form of WFDEI.

A3: We have added more detail on the WFDEI dataset by mentioning that it is based on ERA-Interim data, and additionally adjusted with respect to monthly gridded observations of temperature, precipitation, cloud cover, and atmospheric aerosol loading.

Page 1: Please change "difficulties to correctly capture" to "difficulties in correctly capturing".

A4: Done.

Page 2: Line 6: Please change "This lead" to "This leads".

A5: Done.

Page 2, Line 9: Please change "which become" to "which have become".

A6: Done.

Page 2: Line 33: Please change "helps to better" to "help better".

A7: Done.

Page 2, Line 37: Please change "investigate the role of the land surface model calibration for the skill of" to "investigate the influence of the land surface model calibration on the skill of".

A8: Done.

Page 3 Line 8: Please change "skills" to "skill".

A9: Done.

Page 4, Line 26: Not sure why "sub" needs to be within parenthesis here?

A10: We have removed the brackets.

Page 7: Line 21. I do not see the map in Figure S1.

A11: See answer A2.

4.2.3. Please change "skills" to "skill".

A12: Done.

Advancing land surface model development with satellite-based Earth observations

Rene Orth[1,2], Emanuel Dutra[1,3], Isabel F. Trigo[3,4], and Gianpaolo Balsamo[1]

[1] European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, UK
5   [2] Institute for Atmospheric and Climate Science, ETH Zürich, Universitätstrasse 16, CH-8092 Zürich, Switzerland
[3] Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal
[4] Instituto Português do Mar e da Atmosfera, 1749-077 Lisboa, Portugal

Correspondence to: R. Orth (rene.orth@env.ethz.ch)

10   **Abstract.** The land surface forms an essential part of the climate system. It interacts with the atmosphere through the exchange of water and energy and hence influences weather and climate, as well as their predictability. Correspondingly, the land surface model (LSM) is an essential part of any weather forecasting system. LSMs rely on partly poorly constrained parameters, due to sparse land surface observations. With the use of newly available land surface temperature observations, we show in this study that novel satellite-derived 15 datasets help improve LSM configuration, and hence can contribute to improved weather predictability.

We use the *Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land* (HTESSEL) and validate it comprehensively against an array of Earth observation reference datasets, including the new land surface temperature product. This reveals satisfactory model performance in terms of hydrology, but poor performance in terms of land surface temperature. This is due to inconsistencies of process representations in the model as 20 identified from an analysis of perturbed parameter simulations. We show that HTESSEL can be more robustly calibrated with multiple instead of single reference datasets as this mitigates the impact of the structural inconsistencies. Finally, performing coupled global weather forecasts we find that a more robust calibration of HTESSEL also contributes to improved weather forecast skills.

In summary, new satellite-based Earth observations are shown to enhance the multi-dataset calibration of LSMs, 25 thereby improving the representation of insufficiently captured processes, advancing weather predictability and understanding of climate system feedbacks.

1. Introduction

The land surface forms an essential part of the climate system. It interacts with the atmosphere through the exchange of water and energy and hence influences weather and climate (Seneviratne et al. 2010). Soils, 30 vegetation and water bodies store large amounts of energy and moisture. Through this storage and control capacity, the land surface can accumulate and maintain anomalies induced by the atmospheric forcing (Orth et al. 2013). These persistence characteristics and the associated predictability make the land surface an important potential contributor of weather and climate forecast skill (Orth and Seneviratne 2014, Orth et al. 2016). Furthermore, the land surface can play an important role during extreme events (Mueller and Seneviratne 2013, 35 Miralles et al. 2014, Hauser et al. 2016). For instance dry soils can contribute to the intensification of heat waves but buffer floods, whereas wet soils can mitigate hot extremes but enhance the risk for flood events.

However, state-of-the-art land surface models have difficulties ~~to~~in correctly captur~~e~~ing land surface dynamics and the related coupling with the atmosphere (Beven and Binley 1992, Beven 2001, Wang et al. 2014, Trigo et al. 2015) and show margins for improvement when compared to simple well-tuned models (Best et al. 2015,

Haughton et al. 2016). This is hampering a full exploitation of related predictability, and the accurate representation of extreme events.

The shortcomings of the models are partly related with sparse observations and the spatial heterogeneity of soils and vegetation. Until recently, available observations were not sufficient to satisfactory constrain complex land surface models which include relevant physical processes required to represent the land-atmosphere coupling. This leads to the paradox situation that these complex models could not outperform simple conceptual models with a very simplified representation of processes, as these can be more accurately calibrated with the few available observations (Orth and Seneviratne 2015, Best et al. 2015).

This might change in the coming years thanks to new satellite-derived datasets which have become increasingly available. Related products are already available for essential variables such as surface soil moisture (Liu et et al. 2011, Liu et al. 2012, Wagner et al. 2012) or terrestrial water storage (Swenson and Wahr 2006, Landerer and Swenson 2012). Also information on the heterogeneity of the land surface has strongly improved thanks to satellite-based observations (e.g. global land cover facility, http://glcf.umd.edu, and harmonized world soil database, http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/ [accessed on 11 September 2016]). The unprecedented spatial and temporal coverage of these data offer the potential to enhance the calibration/optimization of unconstrained parameters in land surface models taking into account the variability in soil and vegetation types.

In this study we employ satellite-derived observations of land surface temperatures (LST) which have a high information content on the surface turbulent flux partitioning and on the global surface properties (Mildrexler et al. 2011). Surface temperatures are inferred from emitted infrared radiation at high temporal frequency such that even the diurnal cycle can be observed (Trigo et al. 2011). While the above-mentioned products help to constrain the land water balance, LST products provide complementary information on the land energy balance. Consequently the LST data is expected to bring further constraints to the surface water/energy budgets and improve the land-atmosphere coupling in land surface models. The product considered in this study is based on data from the geostationary Meteosat Second Generation satellite and provides LST information at high temporal and spatial resolution for Europe and Africa. Especially for the latter region, such satellite-based datasets are essential as ground observations are particularly sparse.

Previous studies used LST data from particular days or particular locations to evaluate land surface models (e.g. Wang et al. 2014, Trigo et al. 2015). It is the first objective of this study to comprehensively assess model performance at large spatial scales and with multi-year LST data. Our second objective is to use an increasing number of Earth observation datasets in addition to the LST data to demonstrate that land surface model performance benefits from a comprehensive calibration against a wide range of observational datasets. While they all include characteristic uncertainties and shortcomings, their joint use could help s to better constrain land surface models. Furthermore, by assessing land surface model output with all the employed datasets we can better understand the functioning of the model, identify inconsistencies, and insufficiently represented processes.

Finally, we also investigate the roleinfluence of the land surface model calibration foron the skill of related coupled weather forecasts. This way, we test to which extent an improved representation of land surface processes can propagate into the (modelled) climate system to yield improved predictions.

2. Methodology

In this study we follow the methodology proposed by Orth et al. 2016 (hereafter referred to as O16) regarding the modelling environments and analysis. Sections 2.1-2.4 provide an overview of the model simulations and their analysis, while full details can be found in O16. O16 used these simulations to analyze the sensitivity of the performance of the land surface model and of the weather forecasting system with respect to particular land surface model parameters. In contrast, we will analyze to which extent the simulations capture observed LST and its dynamics, and show that the use of LST data alongside further reference datasets enables a comprehensive and robust land surface model calibration which is also beneficial for weather forecast skills.

2.1 Model description

2.1.1 Land surface model HTESSEL

The ECMWF's land surface model Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL, Balsamo et al. 2011) is an integral component of the ECMWF Integrated Forecasts System (IFS), that is used in the different forecast and data assimilation systems, ranging from deterministic 10-day forecasts to the ensemble seasonal forecasts. The surface model is responsible for providing the atmospheric boundary conditions (heat, moisture and momentum) by simulating the surface water and energy budgets and the temporal evolution of the underlying soil (temperature and moisture), snowpack and vegetation interception.

The surface energy budget is computed in each grid-box independently for different tiles representing different land cover types (e.g. bare ground, high/low vegetation). At each grid cell, only the dominant types of high and low vegetation, respectively, are considered.The surface energy balance is coupled to the underlying soil (or snow) via the skin conductivity, which is currently a single parameter depending on land cover. This is a simplified approach to represent very complex processes such as within-canopy energy exchanges, while it is crucial for the LST computation.

2.1.2 ECMWF ensemble prediction system

The ECMWF ensemble prediction system (Vitart et al. 2008, 2014) is used daily for global forecasts up to the monthly range and it allows to characterise the uncertainty in the meteorological forecast expressed by the spread of the ensemble members (51-forecast realisations in the operational configuration). The spread of the ensemble varies with the difficulty of predicting a given meteorological event, due to the complex evolution of the atmospheric flow and the local climate and seasonal conditions, and it is a highly valuable information on the likelihood of the forecast being accurate. In this study we use 15-member ensemble forecasts.

2.2 Model simulations

Our main objectives are (i) to study the performance of HTESSEL against multi-year LST data covering two continents, and (ii) to analyze the benefits of calibrating HTESSEL against multiple reference datasets, including LST data. For this purpose we employ two sets of simulations with perturbed model parameters. The corresponding uncoupled HTESSEL simulations, and the coupled forecasts with the ensemble prediction system (that includes HTESSEL) are listed in Table 1 and described in this section.

2.2.1 Uncoupled HTESSEL simulations

The use of HTESSEL in an uncoupled, stand-alone setting is computationally inexpensive and allows to perform long-term simulations across the entire European continent. We analyze 50 simulations of HTESSEL with default and perturbed parameters (see Section 2.2.3). The simulations are computed from 1983-2014, and forced with observed meteorological information as those used in the computation of the ERA-Interim/Land dataset (Balsamo et al. 2015). The first six years are used to spin-up the model, and are therefore is not considered in the analysis.

2.2.2 Coupled forecasts

Coupled forecasts with the ECMWF's ensemble prediction system were computed to assess the response of weather forecast skills to different (parameter) configurations of the land surface model. We employ 11 sets of global forecasts with default and perturbed land surface model configurations. The analysis of the forecasts focuses on the European domain used for the uncoupled HTESSEL simulations, and on northern hemispheric summer. This allows us to study impacts of the land-atmosphere coupling in Europe as this is strongest at that time, and to exclude confounding effects from snow and ice on our analysis.

Correspondingly, the forecasts are initialized on eight start dates (1 May, 15 May, 1 June, 15 June, 1 July, 15 July, 1 August, and 15 August) during 2001-2010 and computed until 45 days lead time. Even though weather predictability is low at such long lead times, the land surface may play a role for forecast skills at these lead times given its profound persistence characteristics (Orth and Seneviratne 2012). Each forecast constitutes an ensemble of 15 members which enables us to perform deterministic and probabilistic skill evaluations. Note that as the forecast sets differ with respect to the land surface model configuration, also the initial land conditions for the forecasts may be different. They are taken from the uncoupled HTESSEL simulations with the respective configuration. Consequently, the forecast skill is not only impacted by the altered HTESSEL configurations during the forecasting period, but also by correspondingly different initial land conditions. All further required initial conditions are taken from the ERA-Interim dataset (Dee et al. 2011), and from the ECMWF ocean reanalysis (Balmaseda et al. 2013). This forecast initialization methodology is also used for the ECWMF operational (sub-)seasonal forecasts.

2.2.3 Parameter perturbations

O16 perturbed a set of six poorly constrained parameters which are deemed important for the performance of the HTESSEL model. They are listed in Table 2.

All selected parameters are perturbed at once (Saltelli et al. 2008). For this purpose, multiplicative factors between 0.25-4 (0.5-2 in the case of the soil depth) were applied to the default values of each of the chosen parameters. This range is chosen to still yield meaningful parameter values while allowing some variation in order to study the sensitivity of model performance to the perturbed parameters. The multiplicative factors were determined with a quasi-random sampling approach (Sobol 1967) which allows to efficiently sample the entire parameter space without introducing correlations between the perturbations of the considered parameters. This way, a large sample of perturbed parameter sets was generated, of which 50 parameter sets were chosen by O16 to limit the computation effort for the uncoupled HTESSEL simulations covering Europe. We use the same parameters sets in this study. Out of these 50 parameter sets, 25 were chosen randomly while ensuring that the

resulting multiplication factors applied to particular parameters are not correlated. The remaining 25 parameter sets were selected from corresponding HTESSEL simulations that agreed best with a suite of Earth observations at 6 locations across Europe. For this purpose the parameter sets were ranked in terms of each reference dataset, and then for each particular parameter set the sum of all ranks was computed. The resulting best-ranked 25

5   parameter sets include the default configuration of the model.

As coupled global forecasts are computationally demanding, they were only computed for a subset of 11 out of the 50 sets of perturbed parameters. This subset includes the default configuration, 5 configurations of the randomly chosen parameter sets, and 5 configurations of the best-performing parameter sets. For the selection of 5 (out of 25, or 24 in the case of the best-performing parameter sets as the default parameter set is already

10   considered for computing the forecasts), all possible sets of 5 configurations were tested to choose the configurations with the lowest correlations between the multiplicative factors of the particular parameters.

2.3 Performance measures

The uncoupled HTESSEL simulations and the coupled forecasts are validated against a range of reference datasets (see Section 3.2), using multiple measures of agreement introduced in this section. These different

15   measures have been applied previously by O16. The variety of considered measures allows to make more efficient use of the information contained in the reference data (Vrugt et al. 2003). In particular, we consider:

- Anomaly correlation:

      We subtract the mean seasonal cycle at each grid cell in both the model output and the reference dataset and correlate the resulting anomalies. The mean seasonal cycle is determined from the entire

20   considered time series at each grid cell.

- Bias:

      The bias is derived by subtracting the mean of the reference dataset from the mean of the model output at each grid cell. Only the time period in which reference data is available is considered in this computation.

25   While these measures are used to evaluate the uncoupled HTESSEL simulations and the coupled forecasts, we use another measure for the coupled forecasts only:

- Reliability:

      The reliability measures the ability of ensemble forecasts to accurately capture the occurrence probability of an event. We consider four events which comprise temperature and precipitation

30   anomalies in the lower and upper tercile, respectively. For the assessment of the reliability, all forecasts from grid cells in a particular region are grouped with respect to the forecasted occurrence probability of a particular event. Then the observed frequency of the considered event across all forecast dates in the group is computed and compared with the forecasted occurrence probability. The resulting relationship between all groups of forecasted probabilities and the respective average observed

35   frequencies (reliability diagram, see e.g. Weisheimer and Palmer 2014) can be assessed through a slope of a linear least-squares regression fit (see O16 for details).

All forecast performance measures are computed for particular regions and lead times. In this context we consider the Northern, Central and Southern European regions (as introduced in Seneviratne et al. 2012), and the forecasts are averaged and evaluated for lead times between 1-15 days, 16-30 days, and 31-45 days. Forecast performances for the entire European domain in terms of anomaly correlation, bias, and reliability, respectively, are then determined by (i) ranking the forecasts obtained with the 11 HTESSEL parameter sets in each of the 3 subregions, and then (ii) ranking the sum of the resulting three ranks for each skill measure. This means the HTESSEL parameter set performing best across Europe in terms of a particular skill metric (e.g. temperature bias) must not necessarily be the best in all considered subregions, but has the lowest sum of the ranks from all subregion rankings.

In line with our forecasts that are initialized and computed during late spring and summer, the evaluation of the uncoupled HTESSEL simulations focuses on May-October to exclude impacts of ice and snow on the quality of the reference datasets and on the strength of the land-atmosphere coupling.

2.4 Computation of parameter sensitivities

We assess the performance of the uncoupled HTESSEL model against LST data and further reference datasets, and analyse its sensitivity to variations in particular parameters. In this context all combinations of performance metrics (considered measures of agreement with all employed reference datasets) and perturbed parameters are considered. Sensitivities are computed from the relationship between model performance and underlying multiplicative factors applied to the considered model parameter. A smoothing function (cubic spline function) is fitted to capture this model performance-multiplicative factor relationship (see Figure S1 in O16 for illustration). The sensitivity is then expressed as the fraction of performance variability captured by the smoothing function, i.e. it is calculated by dividing the performance variability captured with the smoothing by the performance variability computed across all involved multiplicative factors.

2.5 Processing of LST data

LST data are new satellite-derived Earth observations which help to better constrain the land energy balance that was so far only captured by the evapotranspiration reference data.

The LST dataset used in this study (see Section 3.2.1) is available at very high spatial (geostationary projection, 3-km at the sub-satellite point) and temporal resolution (15 minutes) across the Meteosat disk (Freitas et al. 2010, Trigo et al. 2011). Here we use hourly fields re-projected onto a regular 0.05°x0.05° grid covering Europe and Africa, which were subsequently processed to obtain mean daily LSTs and the daily LST range at the resolution of the HTESSEL simulations (0.5°x0.5°). We refer to the daily LST range as the difference between the maximum and minimum hourly value of a given day at a particular location (also referred to as daily temperature range). In addition, also the modelled LST data is filtered to exclude (modelled) cloudy days from the comparison. For the data processing we follow several steps:

1. If any 0.05°x0.05° grid cell has more than 2 missing values (i.e. less than 22 hourly values) on a particular day, all data of that day is disregarded. This ensures that any daily values we compute are based on a representative set of at least 22 hourly observations.

2. Any 0.5°x0.5° grid cell is composed of 100 0.05°x0.05° grid cells. If at least 80 of these contain observations, we compute an hourly average across the available (80-100) grid cells.

6

3. From these hourly averages the daily mean LST and daily LST range of the particular 0.5°x0.5° grid cell is computed.

4. The modelled LST data is filtered with respect to the concurrent simulated cloud cover. HTESSEL outputs cloud cover at each grid cell for 3-hourly periods. LST data from a particular grid cell and day are considered if total cloud cover is below 10% in every 3h-period of that day.

Note that these filtering steps are rather strict to guarantee the best possible comparison with the simulations. Some of the filtering steps might be relaxed for other applications, but a detailed evaluation of this filtering is beyond the scope of this study.

3. Data

3.1 Forcing data

As we are aiming to compare the uncoupled HTESSEL simulations against observation-based reference datasets, we use observation-based meteorological forcing to compute all simulations. For this purpose we employ the 'Watch Forcing Data methodology applied to Era-Interim data' WFDEI dataset (Watch Forcing Data methodology applied to Era-Interim dataWFDEI, Weedon et al. 2014)., which is based on bias-corrected ERA-Interim data, and additionally adjusted with respect to monthly gridded observations of temperature, precipitation, cloud cover, and atmospheric aerosol loading.

3.2 Validation data

3.2.1 Validation of uncoupled HTESSEL simulations

In addition to comprehensively evaluating the LST performance of HTESSEL, it is a main objective of this study to analyze and illustrate the value of using an array of Earth observation datasets instead of single datasets to calibrate a land surface model. For this purpose we consider several reference datasets:

- Soil moisture:

  We use data from 11 stations across Europe, which are displayed in Figure S1. Stations are located in Finland (4), Switzerland (5), and Italy (2), and provide therefore data from all relevant European climate regimes. Data are available from different soil depths, and during different time periods, which both vary with respect to the station (see Table S1). For every station, however, there are at least 4 years of data (Figure S1). Aggregating the data from the different depths, we derive a weighted average (with respect to observed depths) to represent soil moisture within the top meter of the soil. The same is done with the HTESSEL data, using the three uppermost soil layers.

- Total terrestrial water storage:

  This quantity is derived from satellite measurements of temporal variations in the Earth's gravity field. The resulting GRACE dataset (Swenson and Wahr 2006, Landerer and Swenson 2012) provides gridded quasi-monthly water storage anomalies, and spans from 2003-2012. We use the release of the Center for Space Research of the University of Texas at Austin. Note the relatively low spatial resolution of about 2°x2° in Europe. These observations are compared with a weighted average (with respect to soil depth) of HTESSEL soil moisture from all model layers.

7

- Evapotranspiration (ET):

  Gridded monthly ET data is used from the LandFlux-EVAL dataset (Mueller et al. 2013). This dataset is a blend of diagnostic and modelled datasets. Whereas the diagnostic datasets are based on (point-scale and satellite) observations, the modelled datasets are obtained by forcing land surface models with observed meteorological forcing. The dataset covers the period 1989-2005 and is provided at a spatial resolution of 1°x1°.

- Streamflow:

  We employ daily streamflow data from over 400 near-natural, small (~10-100 km^2) catchments distributed across Europe from Stahl et al. 2010. Their locations are shown in Figure S1. The dataset spans through 1984-2007, but we only employ data from 1989-2007 to allow sufficient time for the spin-up of the model (Figure S1). These observations are compared with HTESSEL streamflow data from the respective grid cell within which (most of) a particular catchment is located.

- Land surface temperature:

  We use land surface temperature data generated by the Satellite Application Facility on Land Surface Analysis (LSA SAF, Trigo et al. 2011, Freitas et al. 2010), which is based on observations of the Spinning Enhanced Visible and InfraRed Imager (SEVIRI) onboard the Meteosat Second Generation satellite. The gridded LST data is available for 2007-2014. We compare these data with daily skin temperature data from HTESSEL (see Section 2.5). Except for very moist atmospheric conditions, the error of the LST data is below 1K as compared with in-situ ground observations.

All these reference datasets are complementary in terms of spatial coverage and temporal availability. For example, whereas the soil moisture stations represent particular locations, the GRACE and LST data fully cover the European continent (except for cloudy regions in the latter case), but in lower spatial resolution. And while the ET data help to validate HTESSEL's energy balance during early years (1989-2005), the LST data cover the recent years (2007-2014).

Note that for the in-situ soil moisture and GRACE we only consider anomaly correlations to compute the agreement between the reference data and the HTESSEL output. For ET, streamflow, mean daily LST, and daily LST range we additionally consider the bias. This results in a total of 10 validation metrics which we use in our analysis.

3.2.2 Validation of coupled forecasts

We determine the skill of the coupled forecasts against gridded temperature and precipitation observations from the E-OBS dataset version 12 (Haylock et al. 2008). It is based on corresponding station observation from across Europe which are filtered and interpolated to a regular grid using state-of-the-art methods. See Section 2.3 for the employed measures of agreement between forecasted and observed data.

4. Results

4.1 Skin temperature performance of HTESSEL

We perform the most comprehensive large-scale evaluation of LSTs from a land surface model performed so far, covering 8 years and 2 continents. The LST performance of HTESSEL using its default configuration is displayed in Figure 1. There are significant biases in mean skin temperature (overestimated in HTESSEL by more than 5°C in the Arabian Peninsula), and even more in the daily LST range (underestimated in HTESSEL by up to 10°C in southern Europe and southern Africa). These biases clearly exceed the uncertainty range of the LST reference data, indicating a dominant role of model deficiencies. We find strong spatial differences in terms of the performance of the temporal LST dynamics in HTESSEL with lowest correlations in low latitudes. Interestingly, the performance in terms of biases and dynamics do not correspond, we find regions with low biases but low correlations (e.g. Sahel), or regions with strong biases but good representation of observed daily dynamics (e.g. southern Europe). No results can be computed for tropical Africa because the LST observations are not available over very densely vegetation areas as these are frequently covered by clouds.

We furthermore perform this evaluation for land cover classes; for this purpose we only consider grid cells where the respective land cover accounts for more than 80% of the grid cell area. Note that consequently not all areas are included in this analysis as some regions are characterised by mixed vegetation (e.g. Europe). In the lower part of Figure 1 we find that HTESSEL's performance in simulating daily skin temperature range depends on land cover, with better performance over less vegetated areas. This points to shortcomings in the representation of the soil-vegetation energy flux in HTESSEL. However, the dependency of HTESSEL LST performance on vegetation cover is not found in the case of mean skin temperature performance. The area denoted with the dashed rectangles is the European region on which the rest of this study focuses as multiple Earth observations are available there.

In Figure 2, we analyse the sensitivity of HTESSEL's performance with respect to perturbations in selected, poorly constrained model parameters (x-axis). In this context, HTESSEL's average performance across Europe is determined against several reference datasets (y-axis). All parameters influence model performance in some respect, except for runoff depth and maximum interception. While the HTESSEL performance in terms of hydrological datasets (upper part) is sensitive mostly to stomatal resistance, its skin temperature performance (lower part) is especially sensitive to the skin conductivity parameter. The performance in terms of both groups is partly sensitive to shape of soil moisture stress function. An important implication of this is that skin temperature performance can not be improved without impacting the hydrological performance of the model. As in Figure 1, we find an apparent influence of land cover on skin temperature performance, however, with similar parameter sensitivities across the different land covers. This suggests that any improvement of the skin temperature computation in HTESSEL could improve skin temperature independent of land cover.

Adding to the sensitivities determined over Europe, Figure S2 shows the sensitivity of the LST performance of HTESSEL determined over the entire domain displayed in Figure 1. The results are similar. Outside Europe we can also analyse skin temperature performance over bare soils and find similar sensitivities as for the other considered land covers. Generally, the spatially similar sensitivities support the representativeness of European skin temperature results and suggest that improvements of European LSTs would also translate into African LSTs which correspond better with the satellite observations.

4.2 Added value of calibrating HTESSEL against multiple reference datasets

4.2.1 Comparing calibration results against single reference datasets

In this section we analyse if HTESSEL configurations performing well against particular reference datasets also perform well against other reference datasets, i.e. if a parameter set that yields for example good soil moisture performance also yields realistic LSTs. For this purpose we assess the performance of parameter perturbations performing best against particular reference datasets with respect to all other reference datasets in Figure 3. White colors mean that parameter perturbations which perform well against particular reference datasets (x-axis) also perform well against other reference datasets (y-axis). Vice versa, black colors indicate that they do not also perform well against other reference datasets . Note that in the case of a perfect model and perfect observations this plot would be completely white. The many dark colored fields in Figure 3 indicate that the parameter perturbations performing best against particular reference datasets are different, i.e. there is no parameter perturbation that performs best in all respects. This can be explained by (1) equifinality (i.e. many different parameter sets leading to equally well performing model simulations) as there are 25 pre-selected well performing parameter sets among all 50 considered parameter sets, and by (2) inconsistencies within HTESSEL, especially between hydrological and skin temperature-related processes. This is apparent as for example HTESSEL configurations performing well in terms of LSTs yield particularly poor performance in terms of hydrology, and vice versa. These inconsistencies might be partly associated with missing processes in HTESSEL, for example the over-simplification that a single parameter represents the complex energy transfers between the top of the canopy and the underlying soil.

Investigating the role of equifinality we also perform this analysis with the 25 randomly chosen HTESSEL configurations only as displayed in Figure S3. In general, results are robust with respect to the employed set of parameter perturbations as indicated by the comparable patterns in the plots. This indicates a higher importance of inconsistent process representations in the HTESSEL model than of equifinality. Consequently, all the reference datasets considered in this study are needed to constrain HTESSEL, whereas for a perfect model, one dataset would be sufficient.

An analysis of this kind can moreover be used to assess overall model performance which can be measured with the mean rank (grayness) across all tested combinations. Note, however, that the result is influenced by the selection and the quality of reference datasets.

4.2.2 Comparing calibration results against multiple reference datasets

In this section we assess the relative performance of the 50 HTESSEL configurations against multiple metrics, i.e. against several reference datasets using different measures of agreement between the reference and modelled data. In this context, we compute the ranks of all simulations against all considered metrics. Thereafter we calculate for each simulation the sum of the individual ranks obtained against the considered metrics. This sum of ranks is then a measure of overall performance of each simulation, and can be used to rank the overall performance against multiple metrics.

In Figure 4, we test how the best- and worst-ranked parameter sets rank in the case that the validation metric(s) is/are replaced by the same number of other validation metric(s). For a perfect model and perfect observations we would find that the best- and worst-ranked parameter sets in terms of particular validation metrics are also best- and worst-ranked, respectively, when compared against other validation metrics. For HTESSEL we find that for an increasing number of employed metrics, worst-performing parameter sets tend to also perform worse in the case of replaced validation metrics. This is a main result of this study, it means that poorly performing model configurations can be more robustly identified when assessing model performance against multiple validation metrics.

However, this behavior is not found for the best-performing parameter sets. There are two main reasons for this:

1. The poor correspondence of model performance against the considered single validation metrics as shown in Figure 1. The impact of the underlying partly inconsistent process representations within HTESSEL on the results in Figure 4 increases when using more validation metrics. When computing Figure 4 without the 2 validation metrics for which we find the highest average ranks in Figure 1 (runoff bias and LST range bias), the best-performing parameter sets are more robustly identified with increasing number of employed validation metrics (green dashed line). This can be explained as the performance of HTESSEL assessed against the 2 omitted metrics is most inconsistent with the its performance against the remaining metrics. The opposite is found when re-computing figure 4 without the 2 validation metrics which correspond best with the remaining metrics, i.e. for which we find the lowest average ranks (soil moisture anomaly correlation and GRACE anomaly correlation).

   This implies that the better the model performance rankings against individual metrics correspond with each other (i.e. the more white color there is in Figure 3), the fewer metrics are required to robustly identify best- and worst-performing parameter sets. This supports the previously discussed importance of the mean rank in Figure 3, it furthermore provides an indication of the required number of metrics to calibrate the considered land surface model.

2. Out of the 50 considered parameter sets, 25 were pre-selected as they performed particularly well. Hence the 50 parameter sets are therefore not randomly chosen but contain more well-performing configurations than expected by chance. Consequently, the performance of the best parameter sets are more similar than the performance of the worst-performing parameter sets such that for example ranks 1-5 might correspond to very similar performances. When computing the above analysis with only the 25 randomly selected parameter sets we find a more robust identification of well-performing parameter sets with increased validation metrics as shown in Figure S4, in contrast to the results in Figure 4.

4.2.3 Added value of using multiple reference datasets for coupled forecast skills

Adding to the above analyses we finally investigate if a more robust calibration of HTESSEL against multiple datasets yields more accurate weather forecasts. For this purpose we perform a similar analysis as in Figure 4. We test how the best- and worst-ranked parameter sets rank if the (uncoupled) HTESSEL validation metrics are replaced by (coupled) weather forecast skill metrics. The results are displayed in Figure 5.

11

Also in this analysis we find benefits of using multiple validation metrics. The parameter sets ranked best (worst) yield better (worse) forecast performance for an increasing number of employed validation metrics. It is another main finding of this study that land surface model calibration against multiple reference datasets instead of a single reference dataset can lead to better weather forecast performance. This result is found at all considered forecast lead times (1-15 days, 16-30 days, and 31-45 days). However, the differences between few and many considered metrics are smaller compared with Figure 4. This can be explained by (1) fewer tested parameter sets (lower signal-to-noise ratio) and by (2) low predictability of temperature and especially precipitation at long lead times (e.g. 31-45 days) such that there is poor forecast skill whatsoever independent of the HTESSEL configuration. Interestingly, forecast skills expressed as anomaly correlations or biases are more related with land surface model calibration than the reliability skill, as can be seen from the large distance between the dashed red and green lines in Figure 5 compared with that of the dotted lines.

In summary, these results underline the importance of land surface (model calibration) for coupled weather forecast skill (Koster et al. 2011), especially in terms of anomaly correlations and bias, and to a weaker extent also in terms of reliability.

5. Conclusions

In this study we assess the performance of ECMWF's land surface model HTESSEL against comprehensive satellite-based land surface temperature observations. In this novel analysis, we focus on the mean LST bias and the simulated LST temporal dynamics, and find overall unsatisfactory performance. There is no region across Europe and Africa where both the mean LST and the dynamics are well captured by the model. The performance is poorest over high vegetation and improves for low or no vegetation. The particularly poor performance over high vegetation suggests models deficiencies related with the representation of the energy exchanges between the top of the canopy and the underlying soil.

Novel Earth observation data such as the LST dataset add to existing reference datasets, and we furthermore highlight the benefit of employing multiple reference datasets altogether in LSM analysis and calibration. They enable a more robust calibration and can therefore help to address the problem of constraining increasingly complex state-of-the-art LSMs. In this context, we also show that a better constrained LSM also contributes to improved weather forecasts. These results suggest the use of comprehensive objective functions in model calibration. Such a function should be composed of various parts assessing the agreement between the model simulations and a variety of reference datasets, using multiple metrics. This way, model parameters can be adjusted more reliably to yield reasonable model performance in terms of various variables, and to capture possible couplings between them.

Based on the analysis in Figure 4 we can even infer how many metrics (i.e. reference datasets and measures of agreement with it) are sufficient to robustly calibrate HTESSEL. While in the figure we only consider up to 5 metrics (as more can not be replaced in case of a total of 10 metrics), extrapolation of the results towards more metrics indicates that HTESSEL can be robustly calibrated against the 10 metrics used in this study. While this means that poorly performing parameter sets can be identified with the considered reference data, we can, however, not robustly determine best-performing parameter sets. This is due to shortcomings in physical process representations in HTESSEL where for instance the bias of the simulated daily LST range can not be improved without degrading the simulated hydrology. While it is beyond the scope of this study to improve HTESSEL, identifying its shortcomings in representing LSTs and in capturing the corresponding links with land hydrology serves as a valuable basis for future development efforts.

12

Even though the above-described main results of this study should be very relevant for the land surface modelling community, there are caveats in our analysis:

1. The results are valid for the models used (HTESSEL and ECMWF ensemble forecasting system), and the parameters we chose to perturb. Future research is needed to analyze if the methodology and results are transferable to other models.

2. The results are based on the reference datasets and metrics applied here, and on their involved uncertainties. Even though we partly assessed the role of the suite of employed metrics (leaving out 2 metrics at a time), it is not clear if similar findings would be obtained with different reference datasets which inherit different uncertainty characteristics.

3. The assessment of the uncoupled HTESSEL simulations is partly based on the time period considered in the coupled forecasts (2001-2010). This might lead to an overestimation of the benefits of a robustly calibrated land surface model for coupled forecast performance.

4. Our findings might depend on the spatial (0.5°x0.5°, even though this had to be upscaled for the comparison with the ET and GRACE datasets) and temporal resolution (daily) used for the analysis.

Improved constraining of complex LSMs is essential to better exploit their potential and as a basis to represent additional physical processes or updated land-use maps as foreseen in future versions. A more robust model calibration probably also helps to improve the representation of quantities and processes which can not (yet) be constrained with existing observations (e.g. evapotranspiration, sensible heat flux). More physically-based model simulations can also foster improved understanding of (future) climate system functioning, which is particularly important in the context of climate change (IPCC, 2013) as the estimation of climate conditions outside the calibration range of a model is more reliable with physically-based, and therefore complex, models. Finally, as shown in this study, a more robustly calibrated LSM also contributes to improved weather forecasts and is hence valuable for society.

5

References

Balmaseda, M.A., K. Mogensen, A. T. Weaver (2013), Evaluation of the ECMWF ocean reanalysis system ORAS4, Q. J. R. Meteorol. Soc., 139, 1132–1161.

Balsamo, G., S. Boussetta, E. Dutra, A. Beljaars, P. Viterbo, and B. V. den Hurk (2011), Evolution of land surface processes in the IFS, ECMWF Newsletter, 127, 17–22.

Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Muñoz-Sabater, J., Pappenberger, F., de Rosnay, P., Stockdale, T., and Vitart, F. (2015), ERA-Interim/Land: a global land surface reanalysis data set, Hydrol. Earth Syst. Sci., 19, 389–407.

Best, M. J., G. Abramowitz, H. R. Johnson, A. J. Pitman, G. Balsamo, A. Boone, M. Cuntz, B. Decharme, P. A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B. J. J. van den Hurk, G. S. Nearing, B. Pak, C. Peters-Lidard, J. A. Santanello, L. Stevens, and N. Vuichard (2015), The Plumbing of Land Surface Models: Benchmarking Model Performance, J. Hydrometeorol., 16, 1425-1442.

Beven, K. (2001), How far can we go in distributed hydrological modelling?, Hydrol. Earth Syst. Sci., 5, 1–12.

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

Dee, D. P., et al. (2011), The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137 (656), 553–597.

Freitas, S. C., I. F. Trigo, J. M. Bioucas-Dias, F.-M. Göttsche (2010), Quantifying the uncertainty of land surface temperature retrievals from SEVIRI/Meteosat, IEEE Transactions on Geoscience and Remote Sensing, 48(1), 523-534.

Hauser, M., R. Orth, and S. I. Seneviratne (2016), Role of soil moisture vs. recent climate change for heat waves in western Russia. Geophys. Res. Lett., 43.

Haughton, N., G. Abramowitz, A. J. Pitman, D. Or, M. J. Best, H. R. Johnson, G. Balsamo, Boone, M. Cuntz, B. Decharme, P. A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B. J. J. van den Hurk, G. S. Nearing, B. Pak, J. A. Santanello Jr., L. E. Stevens, N. Vuichard (2016), The plumbing of land surface models: is poor performance a result of methodology or data quality?, J. Hydromet., 17, 1705-1723.

Haylock, M. R., N. Hofstra, A. K. Tank, E. Klok, P. Jones, and M. New (2008), A European daily high-resolution gridded dataset of surface temperature and precipitation, J. Geophys. Res., 113, D20,119.

IPCC (2013), The physical science basis, https://www.ipcc.ch/report/ar5/wg1/.

Koster, R. D., et al. (2011), The Second Phase of the Global Land-Atmosphere Coupling Experiment: Soil Moisture Contributions to Subseasonal Forecast Skill, J. Hydrometeorol., 12, 805–822.

Landerer, F. W., and S. C. Swenson (2012), Accuracy of scaled GRACE terrestrial water storage estimates, Water Resour. Res., 48, W04,531.

Liu, Y. Y., R. M. Parinussa, W. A. Dorigo, R. A. M. de Jeu, W. Wagner, A. I. J. M. van Dijk, M. F. McCabe, and J. P. Evans (2011), Developing an improved soil moisture dataset by blending passive and active microwave satellite-based retrievals, Hydrol. Earth Syst. Sci., 15(2), 425–436.

Liu, Y. Y., W. A. Dorigo, R. M. Parinussa, R. A. M. de Jeu, W. Wagner, M. F. McCabe, J. P. Evans, and A. I. J. M. van Dijk (2012), Trend-preserving blending of passive and active microwave soil moisture retrievals, Remote Sens. Environ., 123, 280–297.

Mildrexler, D. J., M. Zhao, and S. W. Running (2011), Satellite finds highest land skin temperatures on Earth, Bull. Amer. Meteor. Soc., 92, 855–860.

Miralles, D. G., A. J. Teuling, C. C. van Heerwaarden, and J. Vila-Guerau de Arellano (2014), Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation, Nature Geoscience, 7(5), 345–349.

Mueller, B., et al. (2013), Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrol. Earth Syst. Sci., 17, 3707–3720.

Orth, R., and S.I. Seneviratne (2012), Analysis of soil moisture memory from observations in Europe. J. Geophys. Res. - Atmospheres, 117, D15115.

15

Orth, R. (2013), Persistence of soil moisture - Controls, associated predictability and implications for land surface climate. PhD Thesis, ETH Zürich.

Orth, R., and S.I. Seneviratne (2014), Using soil moisture forecasts for sub-seasonal summer temperature predictions in Europe. Climate Dynamics, 43 (12), 3403-3418.

5   Orth, R., and S. I. Seneviratne (2015), Introduction of a simple-model-based land surface dataset for Europe, Env. Res. Lett., 10, 044,012, doi:10.1088/1748-9326/10/4/044012.

Orth, R., E. Dutra, and F. Pappenberger (2016), Improving weather predictability by including land-surface model parameter uncertainty. Mon. Weather Rev., 144(4), 1551-1569.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008),
10 Global Sensitivity Analysis: The Primer, Wiley.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling (2010), Investigating soil moisture-climate interactions in a changing climate: A review, Earth-Science Reviews, 99, 125–161.

Seneviratne, S.I., N. Nicholls, D. Easterling, C.M. Goodess, S. Kanae, J. Kossin, Y. Luo, J. Marengo, K.
15 McInnes, M. Rahimi, M. Reichstein, A. Sorteberg, C. Vera, and X. Zhang (2012), In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation 109–230, Cambridge Univ. Press.

Sobol, I. M. (1967), On the distribution of points in a cube and the approximate evaluation of integrals, USSR Computational Mathematics and Mathematical Physics, 7, 86–112.

Stahl, K., H. Hisdal, J. Hannaford, L. M. Tallaksen, H. A. J. van Lanen, E. Sauquet, S. Demuth, M. Fendekova,
20 and J. Jódar (2010), Streamflow trends in Europe: evidence from a dataset of near-natural catchments, Hydrol. Earth Syst. Sci., 14, 2367–2382.

Swenson, S. C., and J. Wahr (2006), Post-processing removal of correlated errors in GRACE data, Geophys. Res. Lett., 33, L08,402.

Trigo, I. F., C. C. DaCamara, P. Viterbo, J.-L. Roujean, F. Olesen, C. Barroso, F. Camacho-de Coca, D. Carrer,
25 S. C. Freitas, J. García-Haro, B. Geiger, F. Gellens-Meulenberghs, N. Ghilain, J. Meliá, L. Pessanha, N. Siljamo and A. Arboleda (2011), The Satellite Application Facility on Land Surface Analysis. Int. J. Remote Sens., 32, 2725-2744.

Trigo, I. F., S. Boussetta, P. Viterbo, G. Balsamo, A. Beljaars, and I. Sandu (2015), Comparison of model land skin temperature with remotely sensed estimates and assessment of surface-atmosphere coupling, J. Geophys.
30 Res. - Atm., 120(23), 12096-12111.

Vitart, F., G. Balsamo, R. Buizza, L. Ferranti, S. Keeley, L. Magnusson, F. Molteni, A. Weisheimer, 2014: Sub-seasonal predictions, ECMWF Tech. Memo. 738, available at: http://old.ecmwf.int/publications/library/do/references/show?id=91226.

Vitart, F., A. Bonet, M. Balmaseda, G. Balsamo, J.-R. Bidlot, R. Buizza, M. Fuentes, A. Hofstadler, F. Molteni,
35 T. Palmer, 2008: The new VarEPS- monthly forecasting system: a first step towards seamless prediction. Q. J. R. Meteorol. Soc., 134, 1789-1799.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resour. Res., 39(8), 2325.

Weedon, G. P., G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo (2014), The WFDEI
40 meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 9.

Weisheimer, A., and T. N. Palmer (2014), On the reliability of seasonal climate forecasts, J. R. Soc. Interface, 11, 20131,162.

Wagner, W., W. A. Dorigo, R. A. M. de Jeu, D. Fernandez, J. Benveniste, E. Haas, and M. Ertl (2012), Fusion of
45 active and passive microwave observations to create an Essential Climate Variable data record on soil moisture, paper presented at SPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS Annals), Volume I–7, XXII ISPRS Congress, pp. 315–321, Melbourne, Australia, 25 Aug.–1 Sept.

Wang, A., M. Barlage, X. Zeng and C. S. Draper (2014), Comparison of land skin temperature from a land model, remote sensing and in situ measurement. J. Geophys. Res., 119, 3093-3106.

16

Table 1: Overview of performed model experiments.

Model	Type	Domain	Spatial resolution	Time period	Number of simulations
HTESSEL	uncoupled	Europe (10°W-50°E, 35°N-70°N)	0.5°x0.5°	1983-2014	50, with different parameter sets
ECMWF ensemble prediction system	coupled forecasts	global	0.7°x0.7°	2001-2010	11, with different parameter sets

Table 2: Summary of perturbed model parameters and their characteristics (adapted from O16).

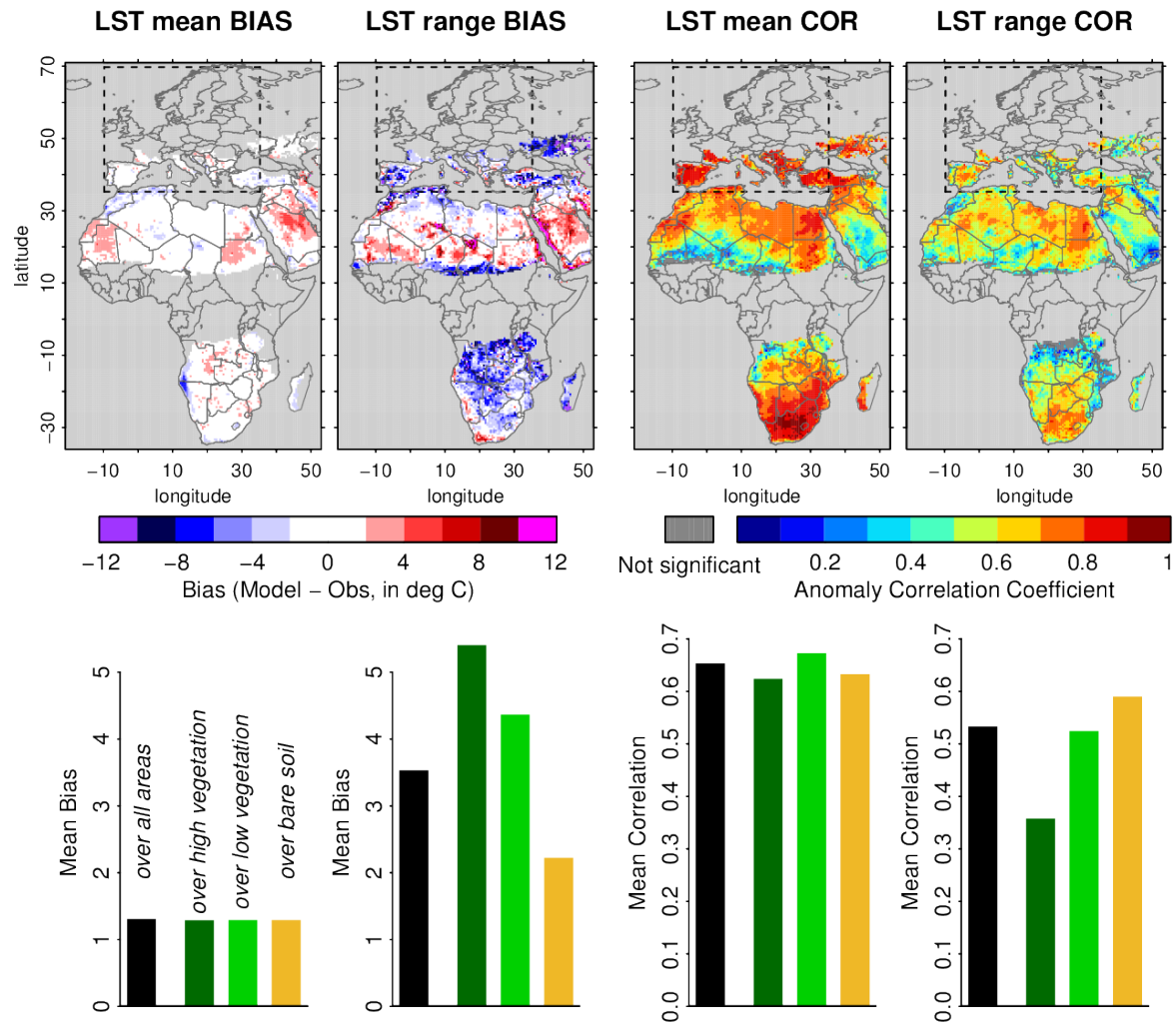| Surface runoff effective depth	Skin conductivity	Minimum stomatal resistance	Maximum interception	Soil moisture stress function	Total Soil depth
Depth over which soil water content and soil water content at saturation are integrated vertically to derive maximum infiltration and eventually surface runoff	Determines coupling of surface energy balance with the underlying surface temperature; dependent on vegetation and stable/unstable conditions	Scales leaf area index in the computation of canopy resistance	Maximum water over a single layer of leaves or bare ground; used to define the interception tile fraction	Determines the shape (e.g. 1 for linear) of dependency of canopy resistance on soil moisture	Lower boundaries of the particular soil layers; top layer not impacted by perturbations to avoid impacts on the fast thermal response

Figure 1: **Evaluation of HTESSEL against LST data. We consider biases (top left) and daily anomaly correlation (top right) of mean daily LSTs and of the daily LST range. Dark gray color indicates correlations which are not significantly different from zero. Light gray color denotes oceans and areas where no data is available. Bar plots (bottom) summarize results for all vegetation types, and for particular vegetation types only. The dashed rectangle in the top plots denotes the European area which most of this study is based on.**
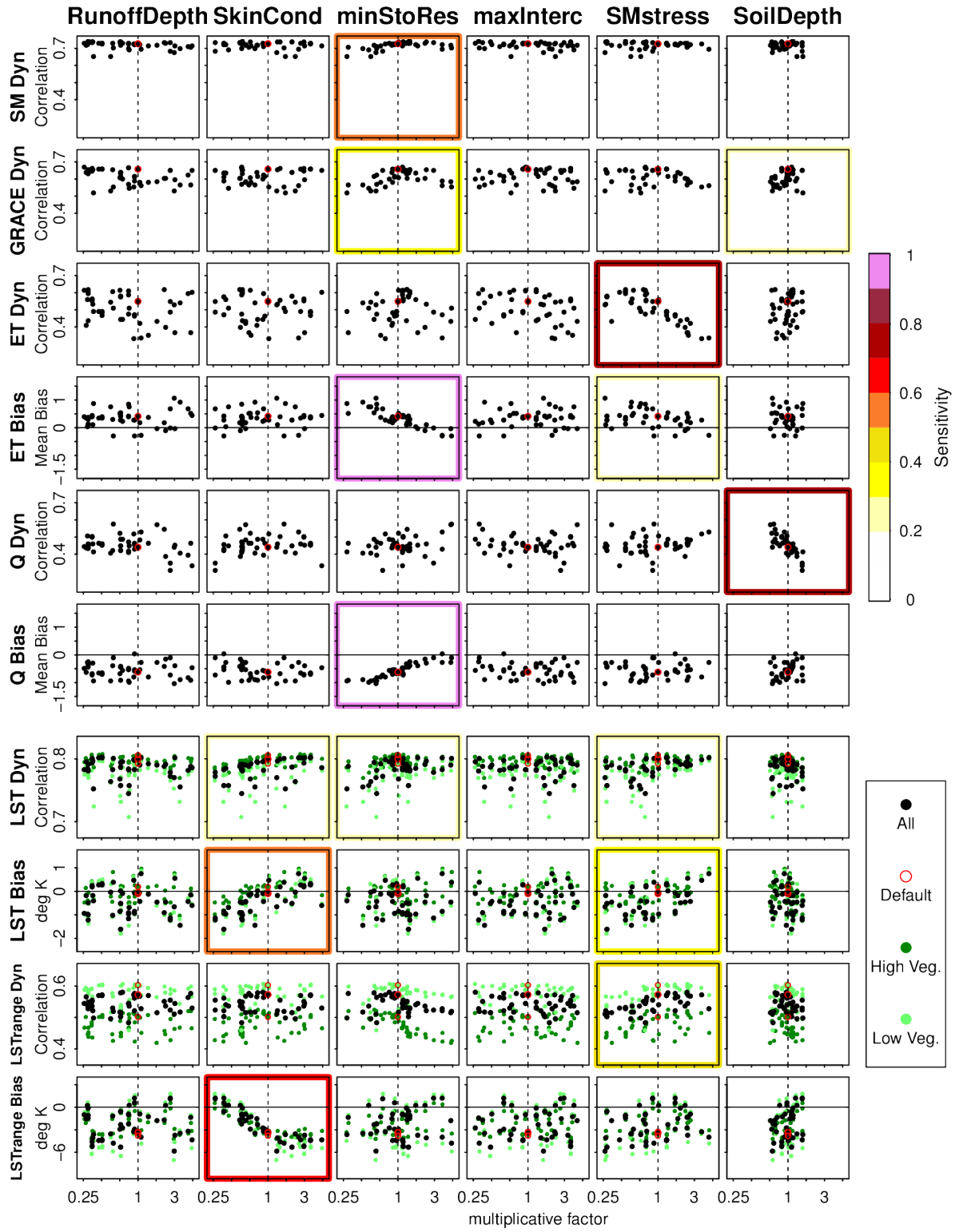
5

Figure 2: Variations of of HTESSEL performance (averaged over the European domain) in terms of anomaly correlations and biases against several reference datasets (y-axis) in response to variations in poorly constrained model parameters (x-axis). The color of each box indicates the sensitivity of the HTESSEL performance in terms of a particular metric against a particular parameter. Red circles indicate results for the default HTESSEL calibration, and green circles denote results for areas with low/high vegetation only.
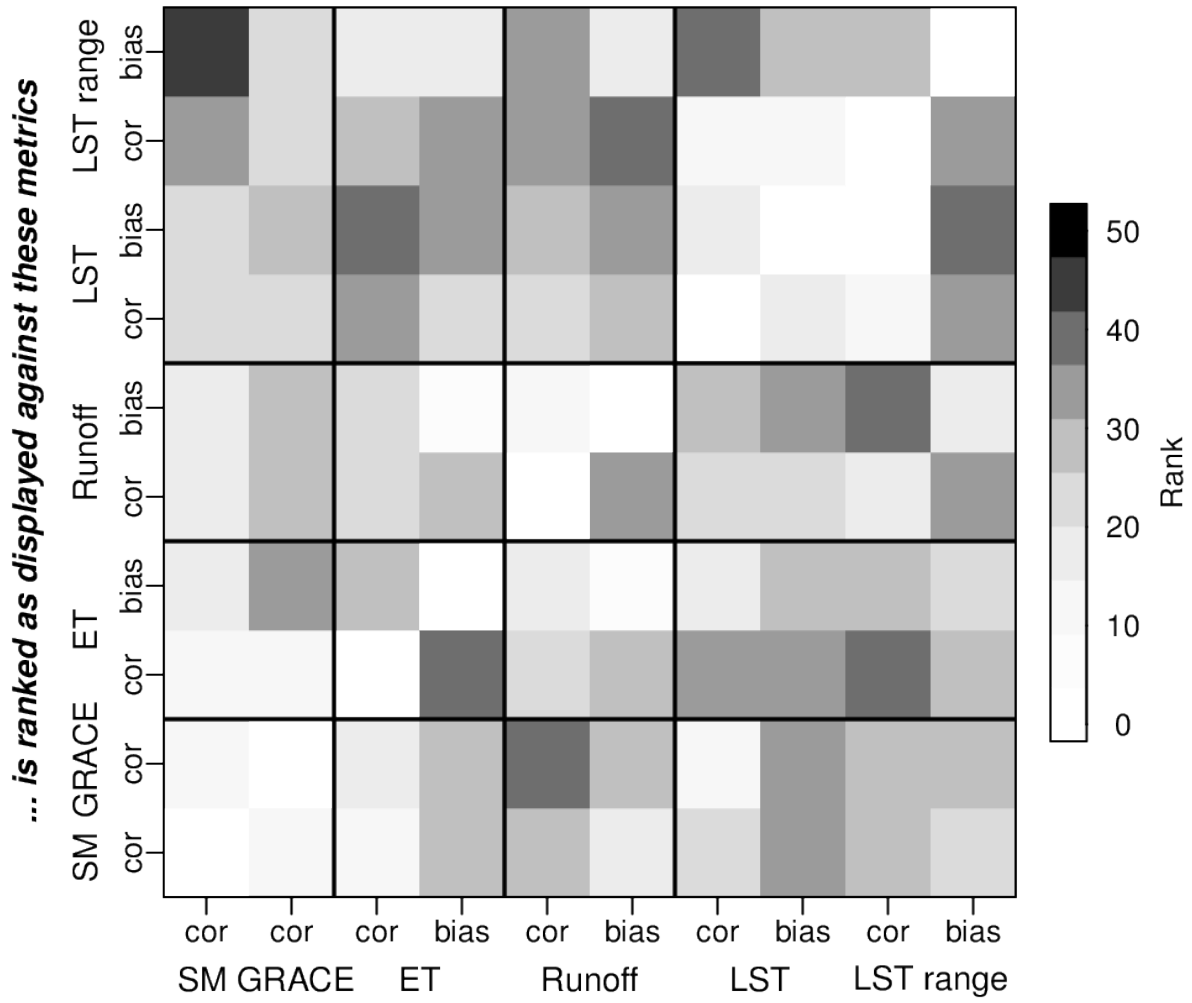
Figure 3: Performance of HTESSEL simulations with the different best-performing parameter sets as assessed against all particular metrics, respectively. For example the HTESSEL simulations with the parameter sets that yield best results in terms of ET bias (see corresponding column) also perform well in terms of runoff bias (white color) but not in terms of ET correlation (dark color).

Figure 4: Comparing the rankings of HTESSEL simulations with all considered parameter sets when replacing a given set of evaluation metrics with an equal number of other metrics. The red curve displays the average rank of the previously worst-ranked parameter sets, and the green curve denotes the average rank of the previously best-ranked parameter sets. For each number of metrics, all possible combinations out of the 10 metrics employed in this study are considered and the mean results are displayed.
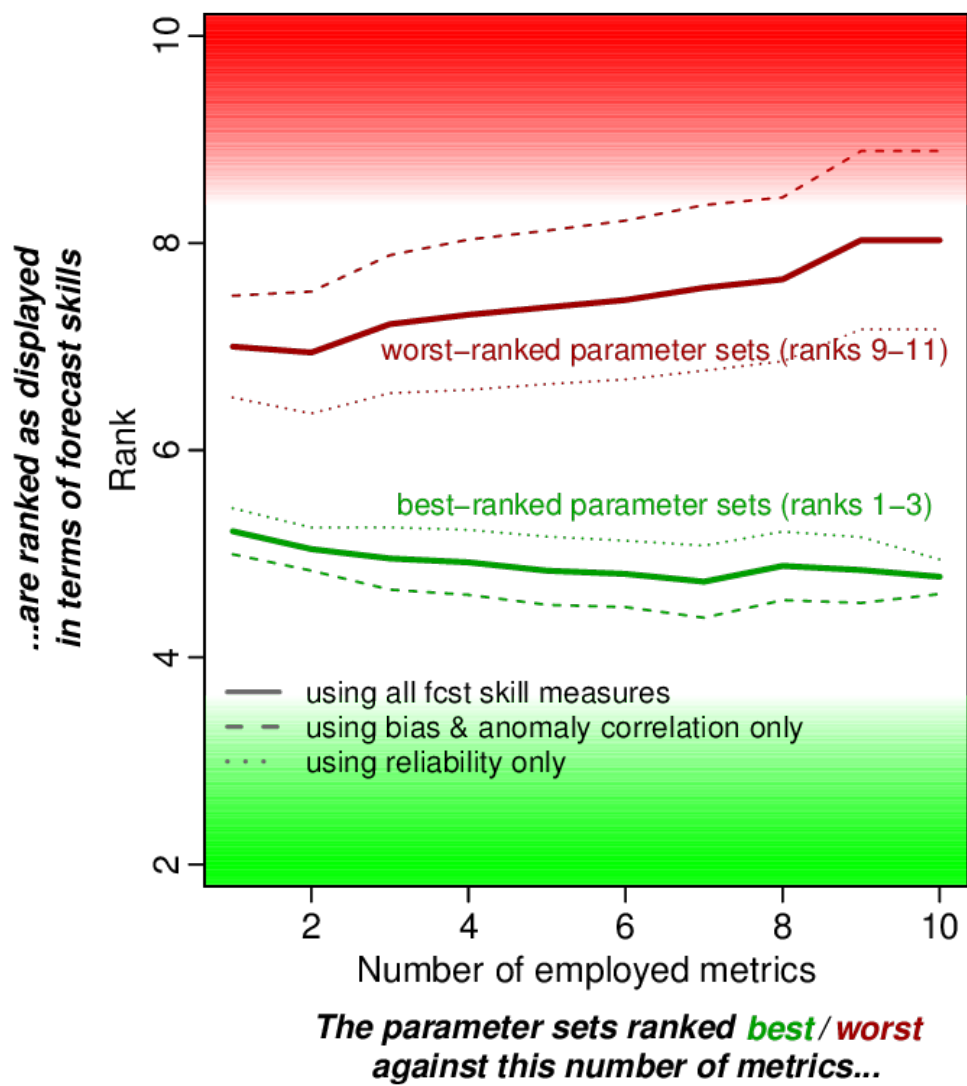
5

Figure 5: Similar to Figure 4 but comparing the performance (ranking) of HTESSEL simulations with 11 parameter sets across uncoupled evaluation (x axis) and coupled forecast skills (y axis).