Improving Weather Predictability by Including Land Surface Model Parameter Uncertainty

RENE ORTH

Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

EMANUEL DUTRA

European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom

FLORIAN PAPPENBERGER

European Centre for Medium-Range Weather Forecasts, Reading, and School of Geographical Sciences, University of Bristol, Bristol, United Kingdom

(Manuscript received 19 August 2015, in final form 27 November 2015)

ABSTRACT

The land surface forms an important component of Earth system models and interacts nonlinearly with other parts such as ocean and atmosphere. To capture the complex and heterogeneous hydrology of the land surface, land surface models include a large number of parameters impacting the coupling to other components of the Earth system model.

Focusing on ECMWF's land surface model Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land (HTESSEL), the authors present in this study a comprehensive parameter sensitivity evaluation using multiple observational datasets in Europe. The authors select six poorly constrained effective parameters (surface runoff effective depth, skin conductivity, minimum stomatal resistance, maximum interception, soil moisture stress function shape, and total soil depth) and explore their sensitivity to model outputs such as soil moisture, evapotranspiration, and runoff using uncoupled simulations and coupled seasonal forecasts. Additionally, the authors investigate the possibility to construct ensembles from the multiple land surface parameters.

In the uncoupled runs the authors find that minimum stomatal resistance and total soil depth have the most influence on model performance. Forecast skill scores are moreover sensitive to the same parameters as HTESSEL performance in the uncoupled analysis. The authors demonstrate the robustness of these findings by comparing multiple best-performing parameter sets and multiple randomly chosen parameter sets. The authors find better temperature and precipitation forecast skill with the best-performing parameter perturbations demonstrating representativeness of model performance across uncoupled (and hence less computationally demanding) and coupled settings.

Finally, the authors construct ensemble forecasts from ensemble members derived with different best-performing parameterizations of HTESSEL. This incorporation of parameter uncertainty in the ensemble generation yields an increase in forecast skill, even beyond the skill of the default system.

1. Introduction

The land surface is an important element in Earth system models (IPCC 2013; Gedney et al. 2014). The hydrology at the land surface is extremely complex; for example, there is lateral and vertical heterogeneity of the soil and the vegetation and nonlinear processes that govern the exchange of energy and water between the

Denotes Open Access content.

Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/MWR-D-15-0283.s1.

Corresponding author address: Rene Orth, Institute for Atmospheric and Climate Science, ETH Zurich, Universitätsstrasse 16, CH-8092 Zurich, Switzerland.
E-mail: rene.orth@env.ethz.ch

DOI: 10.1175/MWR-D-15-0283.1

Publisher's Note: This article was revised on 19 July 2017 to correct an error in the Acknowledgments section when originally published.

land surface and the atmosphere (Seneviratne et al. 2010). Capturing these processes with land surface models is a great challenge (Beven and Binley 1992; Beven 2001). Thanks to the advance in computational resources, their complexity is growing as additional processes are implemented (Oleson et al. 2010; Balsamo et al. 2011).

However, despite an improved physical representation of land surface processes, the models' performance is not necessarily increasing (Orth et al. 2015). This is because the calibration of land surface models becomes increasingly difficult as increased model complexity usually involves more (effective) model parameters and other challenges such as increased uncertainty in the model structure. Furthermore, it is challenging to apply model parameterizations at the required spatial scales (Kauffeldt et al. 2013, 2015). The calibration of model parameters is moreover complicated by the scarcity (and uncertainty) of observations of land surface hydrology and of basic characteristics of the soils and the vegetation.

In this study we focus on the European Centre for Medium-Range Weather Forecasts (ECMWF) land surface model Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land (HTESSEL; Balsamo et al. 2011). To address the uncertain representation of the land surface therein, we adopt a comprehensive approach proposed by Orth and Seneviratne (2015) to calibrate HTESSEL against multiple hydrological observation datasets in Europe. For this purpose we determine a number of poorly constrained parameters and perturb them in many different combinations using a quasi-random variation strategy (Sobol' 1967). Employing observed meteorological forcing, the resulting uncoupled model simulations allow us to identify sensitive parameters and, furthermore, a number of different but equally well-performing sets of parameters.

Usually, the scope of the model calibration defines the methodology. For example, in numerical weather prediction systems the land model should contribute to high forecast skill scores (Cloke et al. 2011; MacLeod et al. 2016). However, model calibration within such a coupled system is difficult and computationally demanding.

Here, we address this problem and investigate the transferability of results from uncoupled model experiments to a coupled setting. An uncoupled setting means running the land surface model driven with (observed) atmospheric forcings without any feedback to the atmosphere. A coupled model run includes running both the land surface model and an atmospheric model that computes the atmospheric forcings taking into account the feedback (e.g., evaporation fluxes) from the land surface. We compare model performance across uncoupled and coupled settings by extending the uncoupled model calibration approach and testing the derived best-performing parameter sets in coupled subseasonal ensemble weather forecasts. Thereby, we analyze to which extent a comprehensive coupled modeling system may benefit from an uncoupled model calibration that is more straightforward and much less computationally demanding.

Going beyond model calibration, in a last step we try to provide a new perspective on land model parameter uncertainty. We analyze potential benefits of considering different but equally well-performing land model parameterizations for the forecast ensemble generation. This is done with coupled model experiments, which are rare in other studies owing to the related computational effort. Using equifinal model parameterizations in this approach allows us to jointly capture uncertainties arising from the model parameterization and from the initial conditions as these change in response to parameter changes. Such an approach can be useful as representing surface variable uncertainties is difficult because their distributions can change over time (Tennant and Beare 2014), and it provides an alternative to model uncertainty assessment through testing different physics schemes (Hacker et al. 2011). The following two sections describe the data and methods used in this study, followed by the results discussion of the uncoupled and coupled simulations, and the last section resumes the main conclusions of this study.

2. Methodology

a. Model description

1) HTESSEL

The ECMWF land surface model HTESSEL (Balsamo et al. 2011) computes the land surface response to atmospheric conditions, simulating the surface water and energy fluxes and the temporal evolution of soil temperature, soil moisture content, vegetation interception, and snowpack conditions.

At the interface with the atmosphere, each grid box is divided into up to six land tiles representing different land covers (e.g., bare ground, high/low vegetation, exposed snow, shaded snow, and interception). The surface energy balance is computed independently for each tile prior to solving the soil and snow mass and energy balance. Therefore, for this explicit formulation, a skin conductivity parameter is required to represent the thermal conductivity between the skin layer and the underlying soil (accounting for processes like radiative transfer within the canopy). The interception reservoir is a thin layer on top of the soil/vegetation, collecting liquid water by the interception of rain a collection of

TABLE 1. Overview of performed model experiments.

Model	Type	Domain	Spatial resolution	Time period	No. of simulations
HTESSEL	Uncoupled	Six representative locations (Fig. 1)	—	1983–2012	2300
HTESSEL	Uncoupled	Europe (35°–70°N, 10°W–50°E)	0.5° × 0.5°	1983–2012	50
ECMWF Ensemble Prediction System	Coupled forecasts	Global	0.7° × 0.7°	2001–10	11

dew, evaporating at the potential rate. The vegetation transpiration is modeled by the canopy resistance that is a function of the available radiation, leaf area index, soil moisture water vapor deficit, and minimum stomatal resistance. The soil is discretized in four layers constant globally (with depths of 7, 21, 72, and 189 cm) for the water and energy transfer. Water leaves the soil column in the bottom layer as free drainage, and at the surface, a variable infiltration rate that accounts for the subgrid variability related to orography is used to compute the surface runoff.

2) ECMWF ENSEMBLE PREDICTION SYSTEM

For the coupled subseasonal forecasting experiments we used a model configuration similar to ECMWF's current operational ensemble prediction system (ENS). The ENS used for operational monthly forecasting is a 51-member ensemble of coupled ocean–atmosphere model integrations up to a lead time of 60 days. The first 10 days are performed with a TL639 resolution (~28 km). After day 10, the resolution is TL319 (~50 km). The 60-day forecasts are performed every Thursday and Monday. The ensemble members differ slightly with respect to their initial conditions and stochastic physics to represent the uncertainties inherent in the operational analyses. In this context, initial conditions refer, for example, to initial soil moisture and pressure fields. Differences in the stochastic physics are generated by random perturbations applied to the tendencies in the atmospheric physics. The initial conditions of the respective ensemble members are produced using the singular vector perturbation method (these include perturbations in the extratropics and perturbations in some tropical areas by targeting tropical cyclones) and also perturbations from the data assimilation. The ocean initial conditions are furthermore perturbed by accounting for wind stress uncertainties in the ocean data assimilation.

In comparison with the operation system, our experiment setup differs in resolution and ensemble size. A lower resolution (T255, ~60 km, constant through the entire forecast period) and reduced ensemble size (15 instead of 51) was necessary to reduce the computational cost of these experiments in order to be able to test several land surface perturbations.

b. Model simulations

Understanding the impact of land model parameter uncertainty on its performance is key to answering the question whether accounting for this parameter uncertainty can improve weather forecasts. For this purpose we need to test a large number of parameter sets in a computationally affordable setting. Selecting a subset of parameter sets, we then test and transfer the initial findings to extended model simulations. This is achieved by analyzing three different configurations (see Table 1 for an overview):

1) To test many different parameter perturbations in HTESSEL, we perform uncoupled simulations at six selected locations from 1983 to 2012. Details of the applied parameter perturbations are provided in section 2c(1). The first year 1983 is included to allow the model to reach an equilibrium state (spinup) and is not considered for the analysis. The locations are chosen to represent the northern, central, and southern European climate; they are displayed in Fig. 1. A time step of 1 h is used.
2) To assess the representativeness of the results from the six locations, we perform uncoupled simulations with the land surface model HTESSEL across Europe from 1983 to 2012. These simulations are carried out with a subset of 50 sets of parameter perturbations [see section 2c(2) for details]. The spinup strategy, model configuration, and time step are the same as above.
3) To study the transferability of the results to coupled forecasts, we select 11 out of the 50 parameter perturbations to perform global ensemble forecasts [see section 2c(2)]. We adopt the study design of the Global Land–Atmosphere Coupling Experiment 2 (GLACE-2; Koster et al. 2010) as they also investigate the impact of the land surface (namely of initial soil moisture) in subseasonal forecasts: We focus on Northern Hemispheric summer with forecasts starting every year between 2001 and 2010 at the beginning and middle of each involved month (i.e., 1 May, 15 May, 1 June, 15 June, 1 July, 15 July,

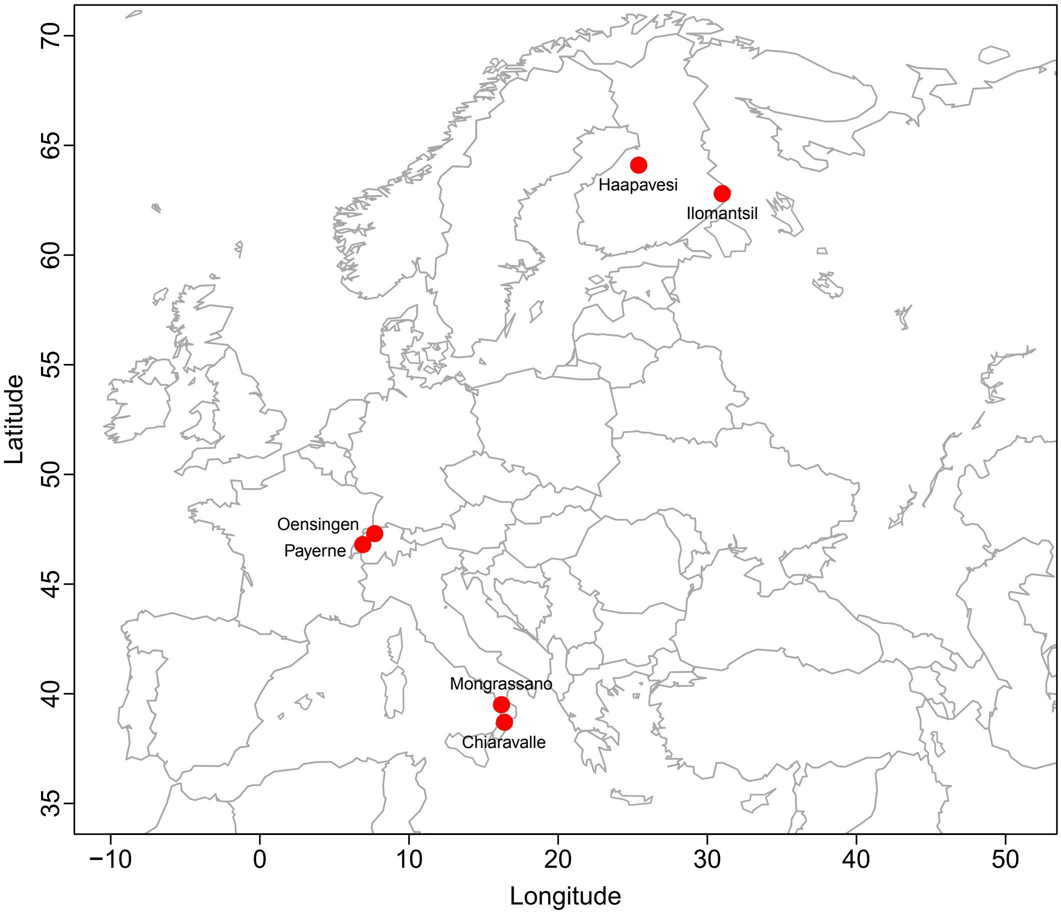

FIG. 1. Selected locations for HTESSEL uncoupled simulations. Furthermore, a subset of 50 HTESSEL simulations is computed across the entire displayed domain.

1 August, and 15 August). The forecasts are computed to a lead time of 45 days and consist of 15 ensemble members, thereby allowing a probabilistic evaluation. The initial atmospheric conditions are taken from the ERA-Interim reanalysis (Dee et al. 2011) for all forecast experiments and the initial land conditions are taken from global uncoupled HTESSEL simulations with the corresponding set of parameter perturbations. That means the initial land initial conditions differ across experiments, while the initial atmospheric conditions are the same (Fig. 2).

c. *Parameter perturbations*

1) SAMPLING THE ENTIRE PARAMETER SPACE

The search for efficient and effective ways to constrain model parameters and their uncertainties is a long-standing topic in land surface modeling, and different methods have been proposed (e.g., Harrison et al. 2012; Beven and Binley 2014).

In contrast to these studies that use Monte Carlo–based techniques, we employ in this study a quasi-random parameter perturbation approach proposed by Sobol (1967) because it allows us to sample the entire multidimensional parameter space more efficiently and it ensures no cross correlations between the perturbations of individual parameters. We have decided not to assume any cross correlations as we feel such an additional assumption is not necessary here because calibrating the model toward optimal performance will properly adjust these correlations anyway. This approach allows extending a purely statistical treatment of the issue with hydrological consideration of error and uncertainty. For a comprehensive sensitivity analysis we perturb a number of selected parameters simultaneously (Saltelli et al. 2008). This allows us to derive the total sensitivity of the model performance against each individual parameter [for applications, see Cloke et al. (2008), Pappenberger et al. (2008), and Dobler and Pappenberger (2013)]. We select six parameters of which we perturb the default values in this study. This

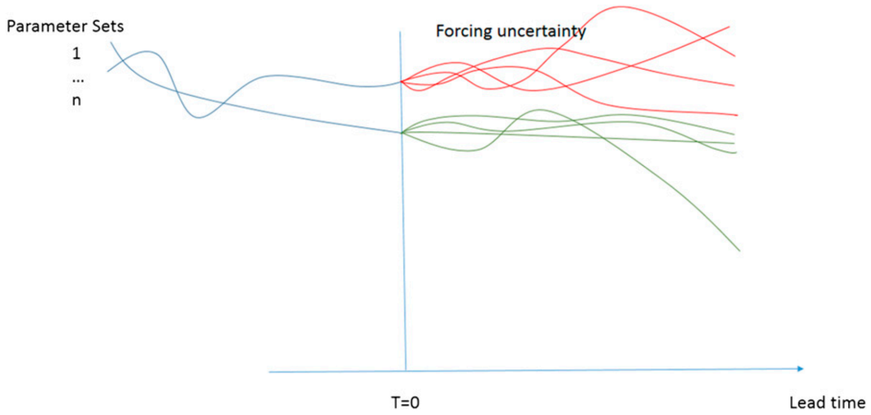

FIG. 2. Illustration of soil moisture evolution before and after forecast start. Initial soil moisture conditions differ, reflecting different model behavior caused by different parameter sets. During the forecast period, the initial soil moisture difference is increasingly overlain by the different atmospheric forcings in the different ensemble members.

choice is based on expert judgment by the model developers and existing literature in this area (e.g., Harrison et al. 2012; Santanello et al. 2013; MacLeod et al. 2016); the parameters are chosen as they are poorly constrained because of scarce measurements and/or large spatial heterogeneity and as they are deemed important for the model's land hydrology. An overview is provided in Table 2. The chosen parameters constitute only a small fraction of the model parameters (i.e., currently of the order of 100). Note that even though more parameters would allow for a more comprehensive model calibration, we are limited by the computational cost as the number of possible combinations of perturbed parameters increases drastically with the number of parameters.

We perturb the parameters by applying multiplicative factors to each of them. The factors are chosen from a range between 0.25 and 4, which keeps the variations within a physically plausible range but still allows for a substantial impact on the model simulations. The multipliers are kept constant in space and time. To further limit the computational effort we use an efficient quasi-random variation strategy to perturb the parameters as described above. To assess interactions between the effects of individual parameters additionally to the total sensitivity of each parameter, we compute a set of 200 perturbations (multiplicative factors) for each of the six parameters. Then, we exchange the first 100 perturbations of the first parameter with the latter 100 perturbations of the same parameter. This is done for each parameter separately and also for every possible combination of 2 parameters, resulting in 200 (original perturbations) + 6 × 100 (exchanged perturbations for each parameter) + 15 × 100 (exchanged perturbations for every set of 2 parameters) = 2300 sets of parameter perturbations. Note, however, that even though we use a large number of parameter perturbations and an efficient sampling strategy, we inevitably miss some regions in the large six-dimensional parameter space, and this consequently impacts our estimation of parameter uncertainties.

2) SELECTING SUBSETS OF PERTURBED PARAMETER SETS

As described above the corresponding 2300 HTESSEL simulations are performed at six locations, whereas a subset of 50 sets of parameter perturbations is chosen for simulations across Europe. One-half of these 50 sets represents the best-performing parameter sets, and the other half is randomly chosen from the remaining 2275 sets to represent the full variability of the parameter space. To determine the performance of the parameter sets we first rank all 2300 simulations by comparing them against each of the four validation datasets [section 3b(1)]. For each parameter set, six ranks are computed based on anomaly correlation for soil moisture, terrestrial water storage, streamflow and ET, and absolute biases for streamflow and ET.

The best-performing sets of parameter perturbations are determined by the following: (i) the sum of the six ranks of the respective simulations must be among the 10% lowest of all 2300 and (ii) the individual ranks must all be lower or equal 1311 (=57% quantile) out of 2300. These numbers were chosen to yield exactly 25 best-performing parameter sets. Note that these include the default parameter set of HTESSEL without any perturbations.

The additional 25 sets of parameter perturbations are selected randomly from the remaining 2275 simulations, thereby ensuring that these 25 parameter perturbations are reasonably different (i.e., mostly uncorrelated). For this

TABLE 2. Summary of perturbed model parameters and their characteristics.

Parameter	Description	Current value	Range of multiplicative perturbation	Reason for poor constraint
Surface runoff effective depth	Depth over which soil water content and soil water content at saturation are integrated vertically to derive maximum infiltration and eventually surface runoff	0.5 m	[0.25, 4]	Effective parameter used to represent subgrid-scale variability that cannot be measured
Skin conductivity	Determines coupling of surface energy balance with the underlying surface temperature; dependent on vegetation and stable/unstable conditions	Vegetation type dependent (7–20 W m^{-2} K^{-1})	[0.25, 4]	Effective parameter that accounts for several processes such as radiative and turbulent transfers between the canopy and the underlying soil
Minimum stomatal resistance	Scales leaf area index in the computation of canopy resistance	Vegetation type dependent (80–250 s m^{-1})	[0.25, 4]	Effective parameter for particular vegetation type
Maximum interception	Maximum water over a single layer of leaves or bare ground; used to define the interception tile fraction	0.2 kg m^{-2}	[0.25, 4]	Spatially variable
Soil moisture stress function	Determines the shape (e.g., 1 for linear) of dependency of canopy resistance on soil moisture	1	[0.25, 4]	This dependence is not directly measured and depends on individual plants physiology
Total soil depth	Lower boundaries of the particular soil layers; top layer not impacted by perturbations to avoid impacts on the fast thermal response	(0.07, 0.21, 0.72, and 1.89 m)	[0.5, 2]	Mean active soil layer that it is not globally constant

purpose we picked a random sample where none of the correlations between the perturbations of any parameter with the perturbations of another parameter exceeds 0.1.

Because of computational constraints, we cannot test all selected 50 parameter sets in coupled global forecasts. Hence we choose a further subset of 11 parameter sets, consisting of the default calibration (see Table 2) and five sets of the (other) best-performing parameters and of the additional parameters, respectively. To make this selection we test all possible combinations of 5 out of 25 parameter sets to select the combination with the most uncorrelated parameter perturbations (see above).

d. *Computing parameter sensitivities*

To assess the sensitivity s of the performance p of a particular output variable of HTESSEL against the perturbations of a particular parameter i we compute (Saltelli et al. 2008)

$$s(p,i)=\frac{\mathrm{var}[E(p\,|\,i)]}{\mathrm{var}(p)}, \qquad (1)$$

where p is the model performance expressed as anomaly correlation or bias and computed against any validation dataset [section 3b(1)], var(p) is the variance thereof, $E(p\,|\,i)$ is the expected value of p conditional on parameter perturbation i, and $E(p\,|\,i)$ is estimated by fitting a cubic spline function [see also Touzani and Busby (2013)], as illustrated in Fig. S1 in the online supplemental material. The sensitivity computed in Eq. (1) is high if there is a relationship between the perturbations of parameter i and the resulting model performance p; that is, if such a fitted spline function can capture a large part of the variability of the individual performance's p. In contrast, the sensitivity is low if the scatter of the performances p is large and random such that a fitted function cannot capture much of the overall variability. The sensitivity varies between 0 and 1; if the performance of any output variable is strongly sensitive to the perturbations of a particular parameter, s will be close to one, whereas otherwise it will be close to zero.

The total sensitivity described above consists of (i) a first-order sensitivity that reflects the direct impact of a particular parameter and (ii) higher-order sensitivity that captures the impact of interactions a particular parameter with other parameters. The experimental design described in section 2b allows us to employ corresponding equations from Saltelli et al. (2009) (Table 2 therein) to compute these components.

e. *Construction of multiparameterization ensemble forecasts*

We test if the consideration of land model parameter uncertainty can improve ensemble weather forecasts. For this purpose we construct ensemble forecasts in three different ways: (i) using ensemble members of all five ensemble forecasts performed with the five best parameter sets [see section 2c(2)], (ii) using ensemble members of all five ensemble forecasts performed with the five additional parameter sets [see section 2c(2)], and (iii) using ensemble members of all three ensemble forecasts performed with the three best-performing parameter sets in terms of temperature and precipitation forecast skills. This ensemble construction is done by randomly selecting individual ensemble members from the forecasts computed with the respective parameter sets. However, aiming for similar initial spread and mean in the resulting constructed ensemble compared with the other ensembles, we ensure that at least 5 out of the 15 members have initial values lower or equal than the unperturbed member of the respective conventional ensemble and at least another 5 should have larger or equal initial values.

f. *Forecast skill measures*

To assess the coupled global forecasts described in section 2b, we infer temperature and precipitation forecast skills [see section 3b(1) for reference datasets]. For a robust evaluation of forecast skill we consider several skill metrics:

1) Anomaly correlation: Correlating forecasted versus observed anomalies. These are derived by subtracting the respective climatological seasonal cycle as determined from the reference data and the forecasts, respectively, over the investigated time period 2001–10.
2) Bias: Here we use the absolute forecasted values and compare them directly with the reference data to obtain the difference.
3) Reliability: To quantify the reliability of the forecasts, we use reliability diagrams [Fig. S2 in the online supplemental material; see also Weisheimer and Palmer (2014)]. These diagrams investigate the ability of the forecasts to detect an event; in this study we consider temperature/precipitation below 33% quantile or above the 66% quantile. Therefore, from all forecasts within a particular region, we select forecasts predicting similar occurrence probabilities (e.g., 10%–20%) of such an event, as determined from their ensemble members; then the mean observed frequency of the event in the selected cases is computed. This is done for all occurrence probabilities that occur in the forecasts. To quantify the relationship between observed frequencies and forecasted probabilities, we consider the slope of a linear least squares regression line using the number of

available forecasts for each range of forecasted occurrence probabilities as weights.

To investigate reliability, we employ climate regions suggested by Seneviratne et al. (2012). For the analysis of forecast skills in Europe, we focus on the respective three regions (north, center, south), but we also average the anomaly correlations and biases across these regions to simplify the analysis.

To evaluate the forecasts, we focus on three ranges of lead times: 1–15, 16–30, and 31–45 days. All forecasts, as well as the reference data are averaged over the respective period such that the skill measures are computed with these averages.

3. Data

a. *Forcing data*

To force the uncoupled HTESSEL simulations (first two rows in Table 1) we use the WFDEI dataset (Weedon et al. 2014), which is derived from bias corrected ERA-Interim data (Harding et al. 2011).

We rerun the uncoupled European simulations with replaced precipitation and radiation data to test the role of the forcing dataset and to ensure that our results are not dominated by errors and uncertainties in the forcings. In this context precipitation and radiation forcings were chosen as uncertainties in those dataset will have the largest impact on modeled land hydrology (Orth and Seneviratne 2015). We replace the bias corrected ERA-Interim precipitation with precipitation from the E-OBS dataset (Haylock et al. 2008), which is derived by upscaling rain gauge measurements. The replacement for radiation is the satellite-derived shortwave and longwave downward radiation data from the CERES experiment (http://ceres.larc.nasa.gov/order_data.php, accessed on 19 June 2015) and the NASA/GEWEX SRB dataset (http://gewex-srb.larc.nasa.gov/, accessed on 19 June 2015). As in Orth and Seneviratne (2015) we join these two datasets after adapting local means and variabilities of SRB to the more reliable CERES data to derive a continuous record from 1984 to 2012. The remaining forcing data (wind, pressure, relative humidity, and temperature) are left unchanged.

b. *Validation data*

1) Validation of uncoupled HTESSEL simulations

We assess the uncoupled model simulations (see previous section) in terms of soil moisture, evapotranspiration, and streamflow. Therefore we use several observation-based meteorological and hydrological datasets as a reference to validate the uncoupled simulations in a multiobjective optimization framework (e.g., Vrugt et al. 2003; Orth and Seneviratne 2015):

(i) Soil moisture measurements from 11 stations across Europe representing different climate regimes. Four stations are located in Finland, 5 in Switzerland, and 2 in southern Italy; for more details on measurement depths and time periods see Table S1 in Orth and Seneviratne (2015). Data from each station cover at least 4 yr and four measurement depths. Locations of two sites in each country are selected to yield the six locations where the first set of HTESSEL simulations is computed (section 2b).

(ii) GRACE terrestrial water storage anomalies derived from satellite measurements (Swenson and Wahr 2006; Landerer and Swenson 2012). We employ the equivalent water thickness data from the current release of the Center for Space Research (The University of Texas at Austin). Even though the spatial resolution of the GRACE data is low (footprint of about 200 km), it allows to validate large-scale patterns of modeled total column soil moisture. The available time period is 2003–12.

(iii) Evapotranspiration (ET) data from the LandFlux-EVAL dataset (Mueller et al. 2013), derived by merging uncoupled land model simulations with diagnostic (satellite based) datasets. This dataset has a spatial resolution of 1° × 1° and covers the time period 1989–2005.

(iv) Streamflow data from >400 near-natural catchments (i.e., with negligible human influence) across Europe (Stahl et al. 2010) from 1984 to 2007.

Note that the validation is performed over different time periods, depending on the reference dataset. Evaluation of the model simulations is performed at the monthly time scale, except for soil moisture where daily observations are available. To validate the first set of simulations at the six locations, we use soil moisture data from the respective locations, GRACE and ET data from the respective grid points, and streamflow data from nearby catchments. However, as there are no data from catchments close to the Italian sites, we do not perform a streamflow validation there.

2) Validation of coupled forecasts

We investigate the performance of the coupled forecasts with respect to temperature and precipitation. To assess their quality in Europe, we use the E-OBS dataset as a reference. As mentioned above, this dataset is derived from a large number of ground observations that are then interpolated to a regular grid (Haylock et al. 2008). To

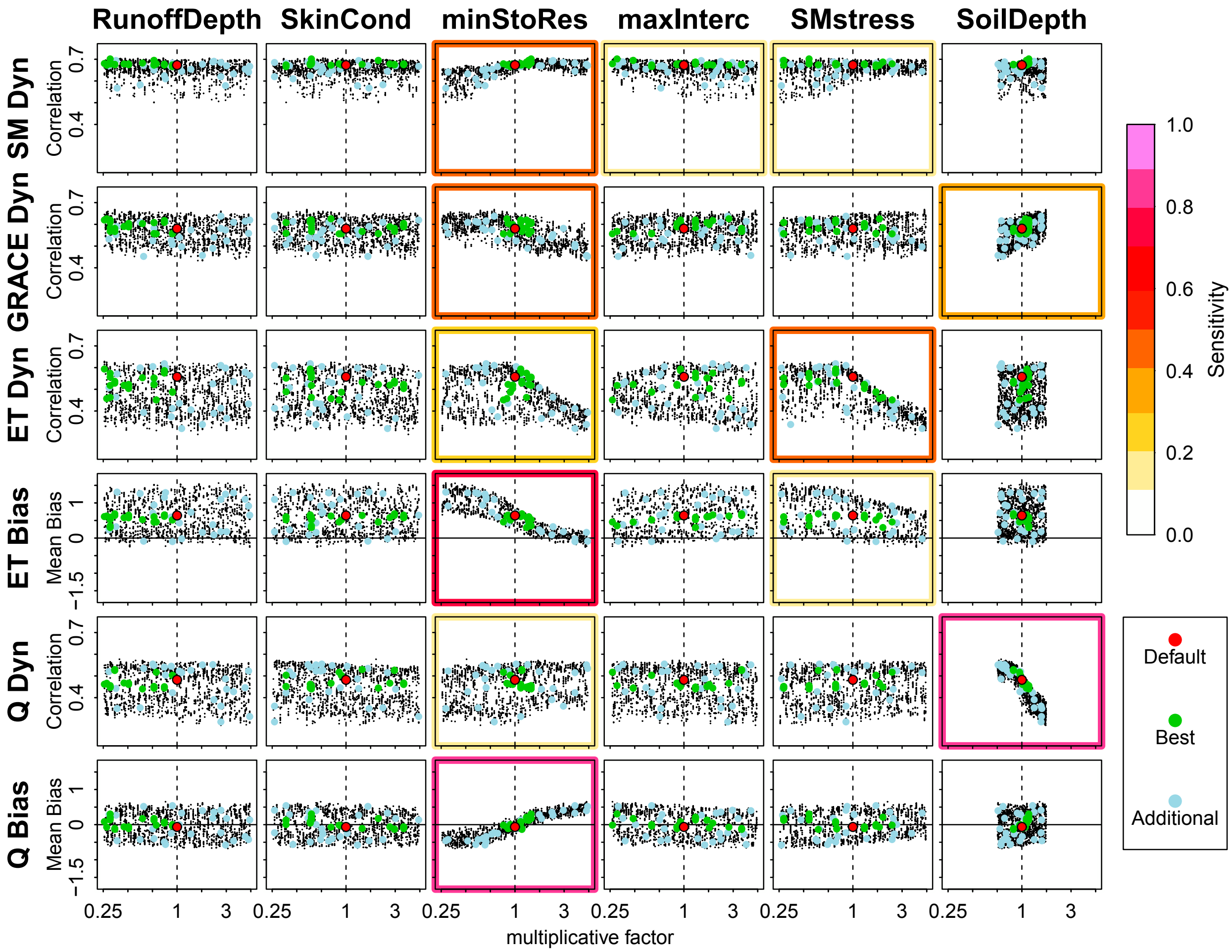

FIG. 3. Performance of uncoupled HTESSEL simulations in terms of soil moisture (dynamics), GRACE terrestrial water storage (dynamics), ET (dynamics, bias), and runoff (dynamics, bias), assessed against multiplicative factors applied to each parameter (i.e., each plot contains 2300 points). Red dots refer to the default HTESSEL calibration, green dots show the other 24 best-performing parameter sets, and blue dots denote the selected 25 additional parameter sets. Sensitivity of any performance metric against any parameter perturbation is expressed by the color of the surrounding box.

infer forecast skills globally we employ data from the ERA-Interim reanalyses, whereby the precipitation data have been corrected with monthly data from the Global Precipitation Climatology Project (GPCP; Adler et al. 2003; http://www.esrl.noaa.gov/psd/data/gridded/data.gpcp.html, accessed on 22 June 2015), as used in the ERA-Interim/Land reanalysis (Balsamo et al. 2015).

4. Results

In this section we first assess the sensitivity of uncoupled HTESSEL simulations to the applied parameter perturbations. We also analyze the role of the forcing and the considered domain in this context. Thereafter, we investigate the sensitivity of forecast skills of coupled subseasonal forecasts to the applied land model parameter perturbations. Finally, we show how accounting for land model parameter uncertainty can improve forecast skills in Europe and around the world, especially during extreme events.

a. Uncoupled simulations

1) PARAMETER SENSITIVITIES AT SELECTED LOCATIONS

We perform uncoupled simulations with HTESSEL at six locations (Fig. 1) using 2300 combinations of six perturbed parameters (sections 2b and 2c; Table 1). The simulated soil moisture, ET, and runoff are evaluated against corresponding observations [section 3b(1)].

The results are shown in Fig. 3. The red point in each plot denotes the default parameterization, which is mostly among the highest correlations or lowest biases, indicating comparatively good performance of the

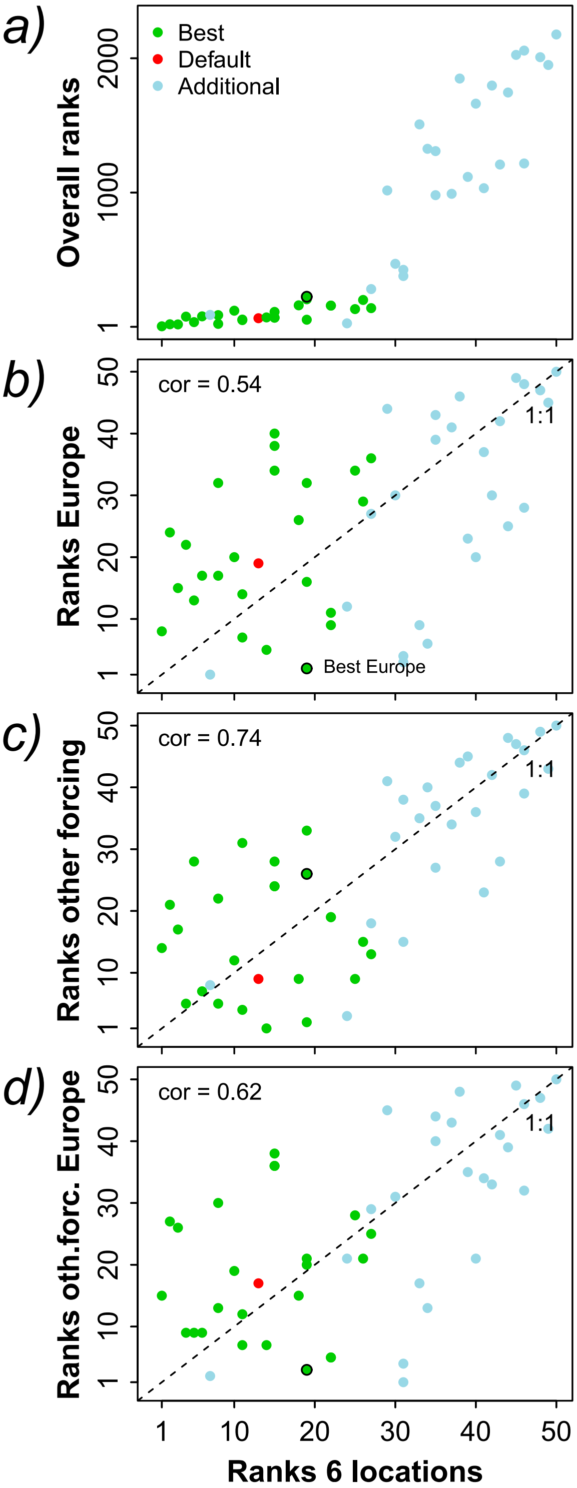

FIG. 4. (a) The overall ranks of the selected subset of 50 parameter perturbations, compared against a ranking among only these 50 parameter sets. (b)–(d) Consistency of ranking of selected subset of 50 parameter perturbations computed at six locations (i) when evaluated across entire Europe, (ii) when simulations are computed with replaced precipitation and radiation forcing, and (iii) when simulations are evaluated across entire Europe and computed with replaced precipitation and radiation forcing.

current model parameterization. As described in section 2c(1), we select 25 best parameter sets and 25 additional parameter sets based on these results. The 25 best parameter sets are shown in green. There is a clustering of their respective multiplicative factors of the minimum stomatal resistance and the soil depth but not of other parameters. This means that these parameters have a profound impact on the model performance, and their respective values need to be within a tight range to ensure good model performance. To sample the entire parameter space with a reduced set of parameter perturbations we select the additional parameter sets [section 2c(1)]. These are shown in light blue and as intended they cover the entire perturbation range of each parameter.

The sensitivities of model performance to the individual parameters (section 2d) are shown by the color of the plot frames. We find a strong sensitivity of model performance with respect to the minimum stomatal resistance. Furthermore, the ET performance is sensitive to the soil moisture stress function, and the runoff dynamics and the total terrestrial water storage are sensitive to the employed soil depth. This can be explained as the soil depth influences the variability of the water storage. Model performance is not sensitive to the three remaining parameters runoff depth, skin conductivity, and maximum interception. However, evaluating HTESSEL's performance in a different context with other than the hydrological observations employed here may yield different results; tests with satellite-based measurements of land surface temperature revealed an important role of the skin conductivity parameter (not shown). In some cases such as for ET dynamics, the sensitivity varies across the range of tested parameter values; for example, the soil moisture stress function has no impact if it is decreased but strong influence if it is larger than the default value. Sometimes results are furthermore contradictory; an increased soil depth, for example, improves the model's performance in terms of terrestrial water storage but deteriorates the simulated runoff dynamics. This underlines the usefulness of our multivariable evaluation approach. Note that with this approach we also validate the model across temporal scales (daily soil moisture, monthly ET, and runoff) and spatial scales [different footprints of the observations; see section 3b(1)], which increases the robustness of the results.

Interestingly, the best parameter sets contain only runoff depth values equal or smaller than the default, even though this parameter is not sensitive (see Fig. S3 in the online supplemental material). Among the best 10% parameter sets (=230) there are also larger runoff depth values; however, only very few of the corresponding parameter sets perform well against runoff and

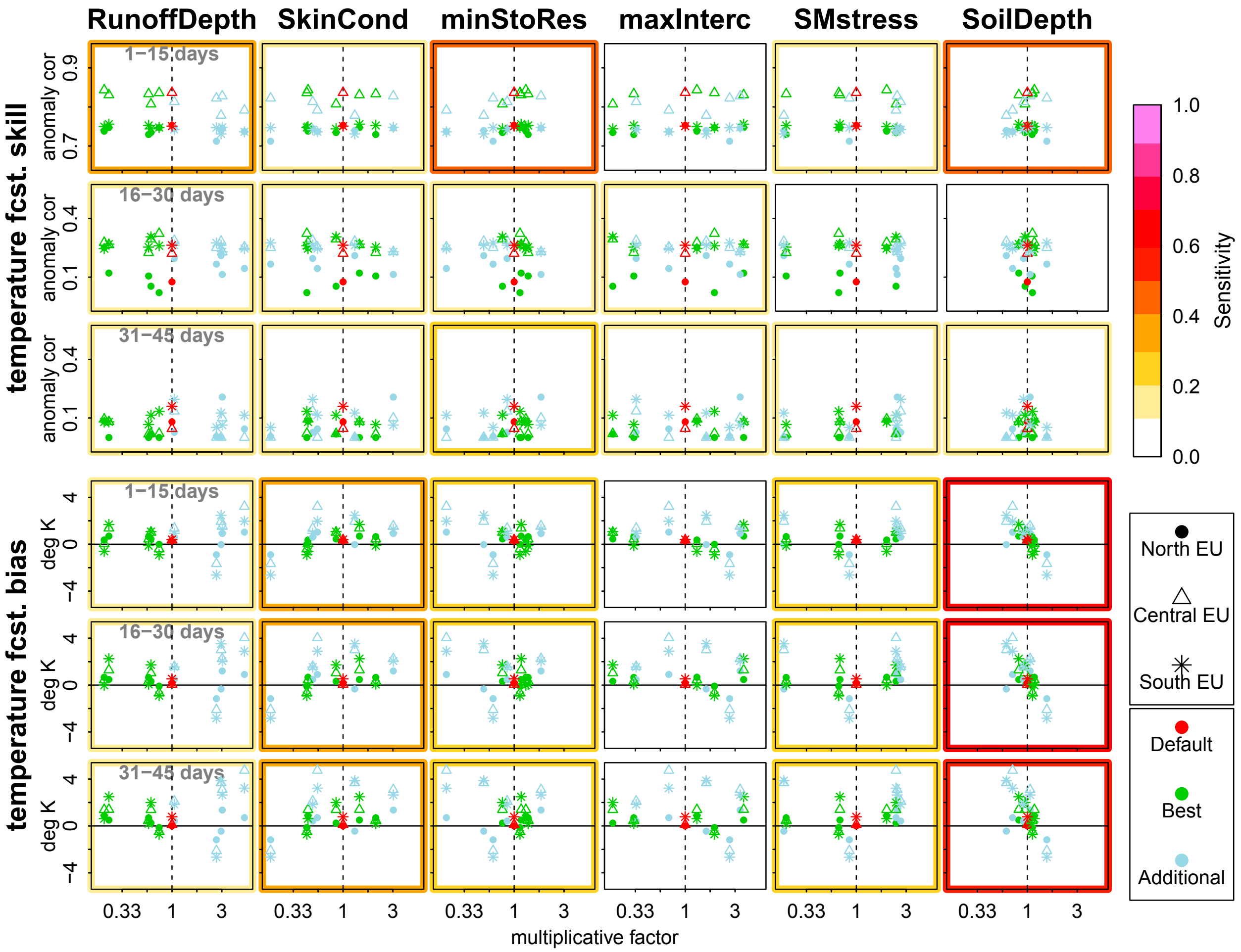

FIG. 5. As in Fig. 3, but for the evaluation of the coupled temperature forecasts using anomaly correlation and bias as performance metrics. Each row refers to a particular lead time. Results are displayed for northern, central, and southern Europe, as denoted by the different symbols.

GRACE dynamics at the same time [applying the 57%-quantile threshold described in section 2c(1)]. None of these finally performs well in terms of ET bias.

2) ROLE OF SPATIAL DOMAIN AND ATMOSPHERIC FORCING

In a next step we test the representativeness of the model performance at the six selected locations. For this purpose we compare ranks of the 50 selected parameter sets (25 best including the default and 25 additional shown in green and light blue, respectively) across different domains and atmospheric forcings. The ranks are inferred from the sum of the ranks in the individual rankings against each observation dataset [section 2c(1)].

Figure 4a shows the overall ranks of the selected 50 parameter sets plotted against the ranks among these 50. It confirms that the 25 additional parameter sets are not only well distributed across the parameter space but also in terms of their performance.

Figure 4b shows that the model performance assessed at the six locations is somewhat representative for the performance across the entire continent, although the rankings differ to some extent. The best datasets still perform better than the randomly chosen parameter sets, even though there is substantial scatter. These results (roughly) validate our approach to start the analysis with only six representative locations to allow testing many parameter sets.

Figure 4c shows the rankings at the six locations with E-OBS and WFDEI atmospheric forcing (section 3a); the results are similar. The performance of any parameter set is hence roughly independent from the employed forcing data; this is a useful characteristic for model calibration in general.

b. *Coupled subseasonal forecasts*

In this section we present results on the forecast skill obtained with different HTESSEL parameter sets

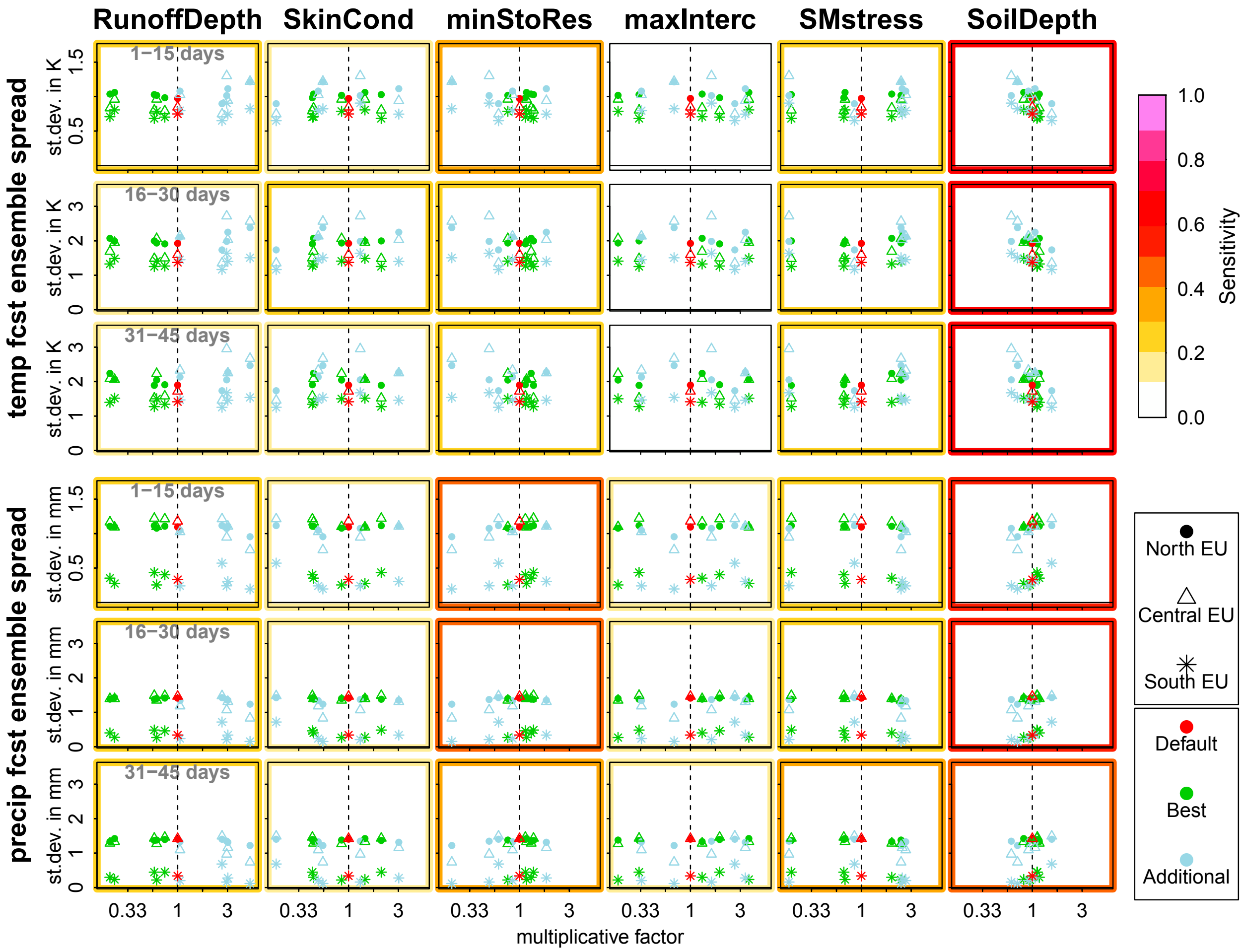

FIG. 6. As in Fig. 5, but focusing on the ensemble spread of the (top) temperature and (bottom) precipitation forecasts.

employed within the coupled forecasting system [section 2a(2)] and its sensitivity toward particular parameters. The focus is mainly on Europe, whereby we distinguish between northern, central, and southern Europe.

1) SENSITIVITY OF FORECAST SKILL ON LAND MODEL PARAMETERS

Figure 5 shows temperature forecasts skills (anomaly correlation and bias) associated with 11 HTESSEL parameter sets at aggregated lead times of 1–15, 16–30, and 46–60 days in the three European subregions. Because of computational constraints we had to select a subset of 11 parameter sets (default + 5 others of the best parameter sets + 5 of the additional parameter sets) from the previously considered 50 parameter sets [section 2c(1)]. Focusing on the obtained sensitivities we find that soil depth and stomatal resistance are impacting forecast skills. Also runoff depth and skin conductivity have some influence. The forecast skills differ partly substantially between the considered subregions. Overall the results are rather noisy as we only test 11 parameter sets. The sensitivities of the anomaly correlations fade away with increasing lead time, whereas the sensitivities associated with bias are rather constant with lead time. This might have to do with the fact that the different land surface model parameterizations impact both the initial land conditions and the land–atmosphere coupling during the forecasting period. Whereas the bias is probably more impacted by the second, the anomaly correlation is influenced by both. The sensitivities related to the anomaly correlation skill probably decrease with lead time because of the increasing uncertainty in the forecast—that is, the reduction of predictability and the decline of the signal-to-noise ratio. The land–atmosphere coupling is somewhat sensitive to most tested parameters with the strongest response to the soil moisture stress function and the soil depth (see Fig. S4 in the online supplemental material). In contrast, the initial soil moisture is sensitive to minimum stomatal resistance

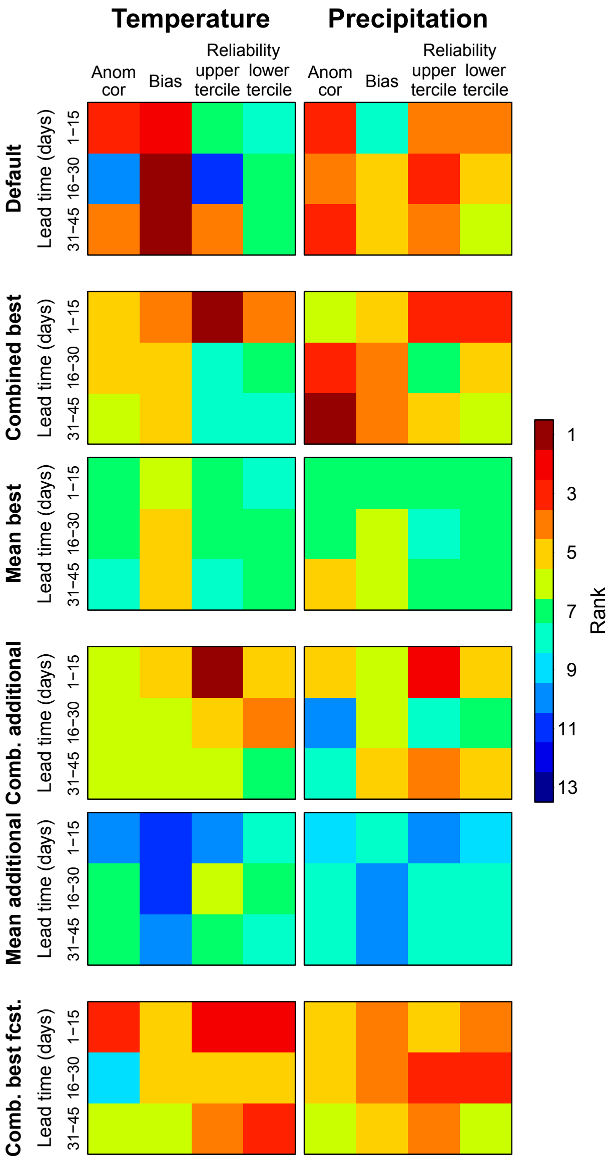

FIG. 7. Ranking of forecasts derived with the selected 11 HTESSEL parameter sets and with three combinations of HTESSEL parameter sets. Ranks are shown for all lead times and all considered skill metrics. (top) Ranks derived with default HTESSEL calibration. (second and third from top) Ranks derived with five best parameter sets. (second and third from bottom) Ranks derived with five additional parameter sets. (bottom) Ranks of forecasts derived with the three best-performing HTESSEL configurations as evaluated from the forecast performance.

and the soil depth (see Fig. S5 in the online supplemental material). The soil depth furthermore impacts the soil moisture spread of the forecast ensemble.

As for the anomaly correlation, we found no clear signal when evaluating the reliability of the temperature forecasts for the lower and upper terciles (see Fig. S6 in the online supplemental material).

Similar to the temperature forecast results, the precipitation forecasts' (see Figs. S7 and S8 in the online supplemental material) sensitivities are rather constant with lead time for the bias but decrease with lead time for anomaly correlation and reliability. Stomatal resistance and soil depth parameters are overall most influential, and the soil moisture stress has some impact on the bias. Interestingly, we find similar influential parameters for temperature and precipitation forecast and, furthermore, these parameters were found to be important in the uncoupled simulations. This means that computationally less demanding uncoupled simulations can help—to some extent—to identify and calibrate key parameters of the land surface mode.

Comparing the green versus the red symbols in the bias plots of temperature and precipitation shows that the best parameter sets tend to make the coupled model drier (especially in central Europe) and warmer (mostly in southern Europe). As a consequence, the land–atmosphere coupling strengthens in central Europe.

In line with the importance of the soil depth for the forecast skills, we find a strong impact of this parameter on the spread of the temperature and precipitation forecast ensembles at all lead times, as shown in Fig. 6. The deeper the soil, the larger is the precipitation ensemble spread and the lower is the temperature ensemble spread. As the land surface water reservoir grows with larger soil depth, an enhanced moisture recycling tends to increase precipitation. At the same time, more moisture dampens temperature variability.

2) IMPROVING SUBSEASONAL PREDICTABILITY BY INCLUDING LAND MODEL PARAMETER UNCERTAINTY?

Moving beyond sensitivities, here we compare forecast skills derived with the different HTESSEL parameterizations. In this context we consider the skills of forecasts derived with the 11 parameter sets. Additionally, we consider artificially constructed ensemble forecasts based on the five best, or the five additional parameter sets that therefore include land model parameter uncertainty [section 2c(2)].

For each lead time and considered forecast skill metric we compute a ranking among the 11 + 3 = 14 considered forecast configurations; the results are shown in Fig. 7. The default parameter set performs comparatively well as indicated by the reddish colors denoting low ranks. This confirms results of the uncoupled simulations in Fig. 3. Only in terms of temperature reliability does the default parameterization perform less well. Comparing the mean skills of the five best and the

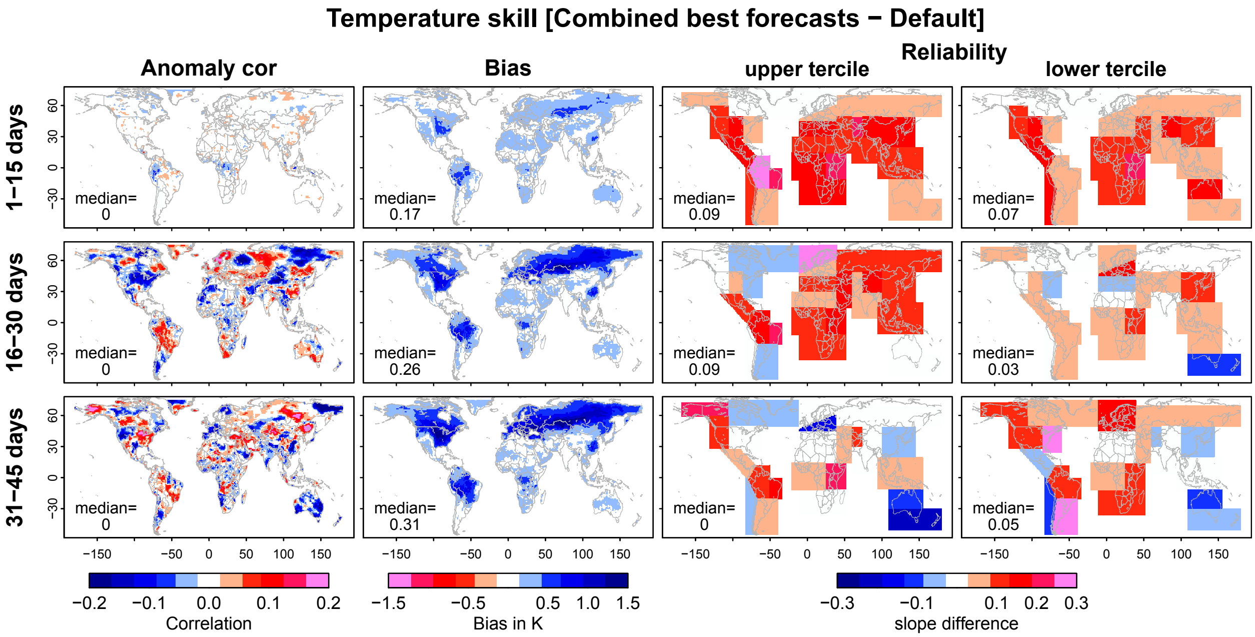

FIG. 8. Global comparison of temperature performance of combined best forecasts vs forecasts with default HTESSEL calibration. Red colors refer to improvements over the default; blue colors refer to a degradation.

five additional parameter sets, respectively, we find better skill across most metrics and lead times in the forecasts using the best parameter sets. This is an important finding, as it illustrates that uncoupled model calibration (against hydrological observations) can improve coupled (temperature and precipitation) forecasts. This is furthermore in line with the similar sensitivities found in the previous subsection. Another main result of this study is that the construction of forecast ensembles from members using different HTESSEL parameterizations (that differ in terms of the six parameters considered here) improves the forecast skill, both for combining the best and additional parameter sets, respectively. Finally, unlike using the best parameter sets determined in the uncoupled simulations, we construct ensembles with the three best parameters sets in terms of forecast skills. As shown in the bottom row, this "combined best" forecast indeed outperforms all other described forecasts, except for the default. It ranks better than the default only in terms of temperature reliability. It seems that three is the minimum amount of required parameter sets within the constructed ensemble; tests with less showed degraded forecast skill whereas tests with more did not yield significant improvements.

With these results, we can answer the question posed in the title of this subsection with *yes*.

However, the current default parameterization performs really well such that it is not easy to apply these findings to improve the operational forecasting system.

3) Going global

Whereas the previous analyses focused on Europe, we present here a global comparison between the combined best forecasts and the default. The results are shown in Figs. 8 and 9 for all considered lead times and skill metrics. Note that reliability slopes are computed over climate regions while the other metrics are computed at each grid point (see section 2f). Confirming results of the previous subsection, we find clearly improved temperature reliability in Europe but also across most other regions of the globe. Only at longer lead times the signal is weaker because the corresponding slopes are small. In terms of the anomaly correlation both forecasts overall perform equally well. We find patchy patterns; with increasing lead time they increase in amplitude and vary spatially. The size of the regions where one forecast is better than the other is about equal. Also for precipitation there is no clear difference in the anomaly correlations, and furthermore the reliability of the two considered forecasts is not systematically different.

Temperature and precipitation biases are globally larger in the combined best forecasts. The biases accumulate with lead time, reaching 1 K and 0.5 mm day^{-1} in many regions. These are remarkable changes in the mean model climate that highlight the role of the land surface parameterization in the coupled system. Note,

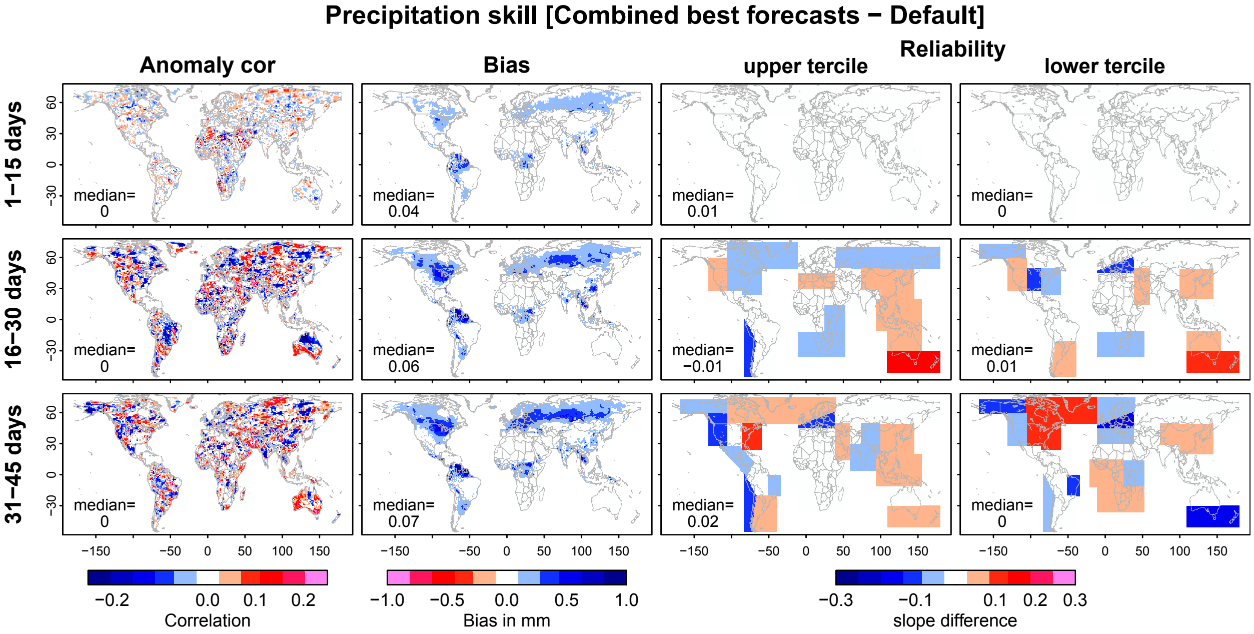

FIG. 9. As in Fig. 8, but for precipitation.

however, that the default model configuration is "tuned" for a reduced bias. This means that parameters in all components of the model (land surface, boundary layer, clouds, radiation, etc.) are adapted accordingly during model development. Following the same methodology with the combined best forecasts may allow us to reduce the biases, even though the improved temperature reliability might be diminished at the same time. A way forward in this context might be to "tune" the coupled system taking into account more skill metrics than only the bias.

4) EXAMPLE FORECASTS OF 2010 RUSSIAN HEAT WAVE

In addition to comparing the combined best forecasts with the default in terms of their mean skill we focus here on extreme events. In this context we choose the heat waves in western Russia in 2010 (Barriopedro et al. 2011) and in central Europe in 2003 (MacLeod et al. 2016) as examples.

The respective forecasts of the onset of the Russian heat wave in the first half of June 2010 are presented in Fig. 10. We compare forecasts of temperature, precipitation, and atmospheric circulation expressed through geopotential height. Forecasts initialized at 1 May do not show any large-scale temperature or precipitation anomaly in the first half of June. Forecasts initialized at 16 May differ between the default system and the combined best forecasts. The latter start to pick up the observed large-scale anomalies in the correct regions, also thanks to the roughly matching forecasted circulation pattern. In contrast, the default configuration does not (yet) forecast a heat wave. This is another main result of this study: the land model parameterization impacts weather beyond the near surface: it can influence the large-scale atmospheric circulation and the associated forecast skill. Forecasts initialized at 1 June capture the upcoming heat wave and the associated large-scale circulation with both configurations.

Whereas the combined forecasts methodology helps an earlier capture of the Russian heat wave, there is no such difference for the onset of the 2003 heat wave (see Fig. S9 in the online supplemental material). The improvement of extreme event forecasts therefore does not necessarily seem to be a general characteristic of the inclusion of land model parameter uncertainty.

Probably the contrasting results for the two considered heat waves are due to different initial (dry) soil moisture anomalies. Whereas soils were not dry in central Europe at the beginning of the 2003 summer such that the event was mainly driven by the large-scale circulation, there was an initial anomaly in western Russia in 2010 (Whan et al. 2015; Prudhomme et al. 2016) that amplified the role of the land–atmosphere coupling.

5. Conclusions

In this study, we have demonstrated how the land model calibration(s) can improve a state-of-the-art subseasonal forecasting system. For this purpose we

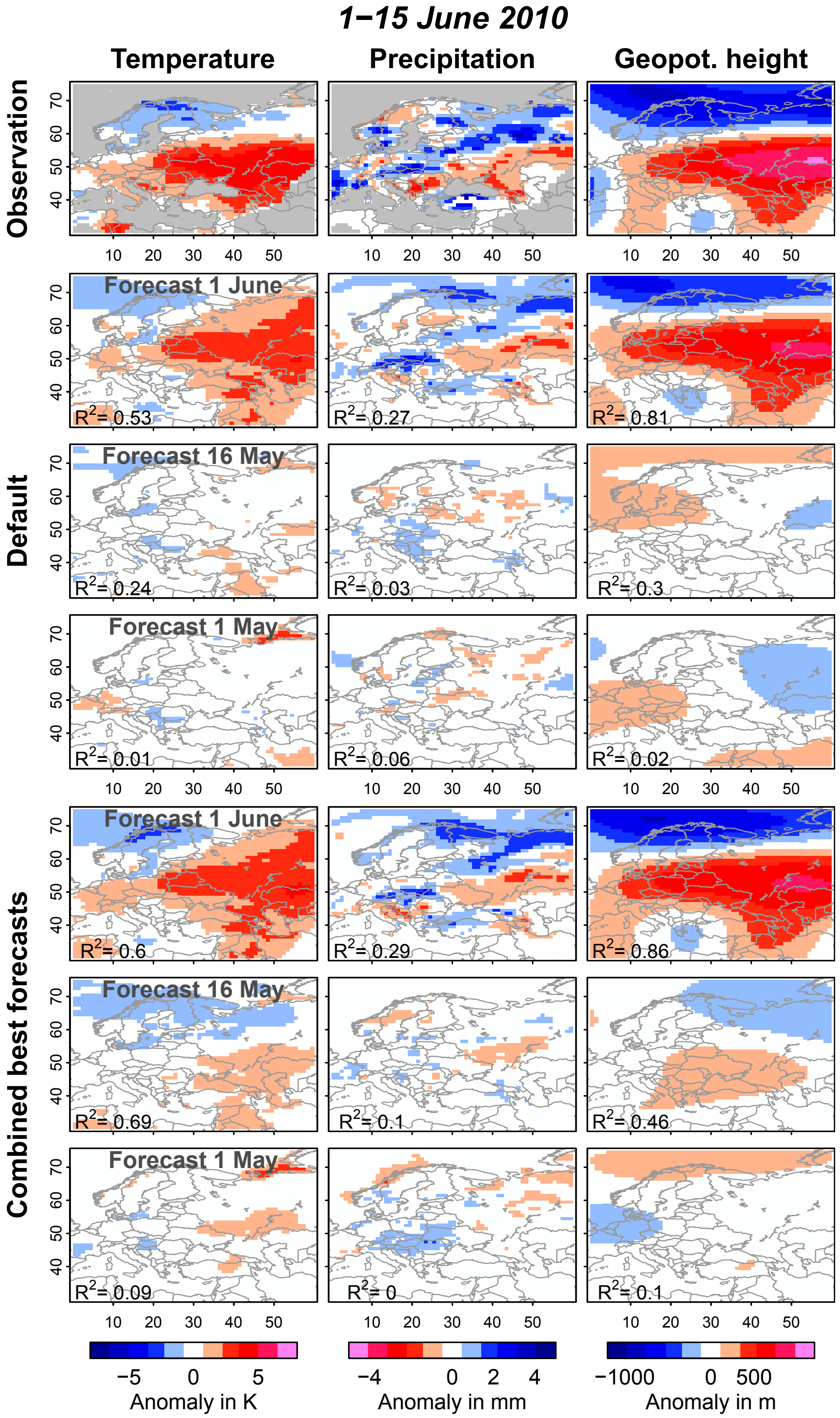

FIG. 10. (second–fourth rows) Comparison of default forecasts against forecasts with (fifth–seventh rows) combined best HTESSEL parameter sets of the onset of the Russian heatwave in 2010. (left) Temperature, (middle) precipitation, and (right) geopotential height at 500 hPa. Rows refer to forecasts initialized at different dates (i.e., where different lead time is considered here). (top) Corresponding observations are shown.

performed uncoupled and hence computationally inexpensive simulations with HTESSEL, which we evaluated against multiple reference datasets to infer several best-performing sets of previously poorly constrained parameters. Employing these sets in coupled forecasts, we illustrated potential benefits of incorporating different but equally well-performing land model parameter sets within the generation of ensemble forecasts. This provides a new perspective on land model parameter uncertainty. Moreover, our analyses highlighted that uncoupled model calibration allows us to improve a coupled forecasting system and, hence, offers a computationally less demanding option to advance a coupled model. Our results underline the importance of the land surface within a coupled Earth system, even though the improvements of the coupled system only occur occasionally as the role of the land surface furthermore varies dependent on conditions such as soil moisture content and the state of the vegetation (Koster et al. 2011). Hence, different results are also found when considering forecasts of mean versus extreme conditions (Fujita et al. 2007) or of dry versus wet conditions (Santanello et al. 2013). Note that the time periods considered in the uncoupled HTESSEL calibration [depending on reference dataset; see section 3b(1)] partly include the time period where the coupled forecasts are validated (2001–10). However, the comprehensive calibration against multiple parameter sets in different time periods should ensure the usefulness of this methodology also for future forecasts. Furthermore, different variables are considered in the uncoupled calibration (soil moisture, evapotranspiration, and streamflow) and the evaluation of the coupled forecasts (temperature and precipitation). Whereas land model parameters are perturbed in this analysis and other studies (Fujita et al. 2007; Harrison et al. 2012; Santanello et al. 2013; MacLeod et al. 2016), which then impacts land surface variables such as soil moisture, similar results can be obtained when perturbing land surface variables directly as in other land surface uncertainty studies (Harrison et al. 2012; Lavaysse et al. 2013; Bouttier et al. 2016). Also, remote sensing data has been used in this context to better constrain land surface parameters (Harrison et al. 2012). All these studies confirm that perturbations in surface hydrology can influence and improve ensemble weather forecasts. These results furthermore indicate that also other than the six parameters we investigated here might have profound impacts on weather forecast skills.

We investigated the role of six poorly constrained parameters and found that the minimum stomatal resistance and the soil depth are most important. Whereas the importance of the stomatal resistance is expected as it directly impacts the exchange of water and energy between the surface and the atmosphere, the important role of the soil depth and hence the water holding capacity has not been previously reported. We find that different model parameters have the strongest impacts on precipitation or temperature. Also, Lavaysse et al. (2013) and Fujita et al. (2007) found that particular land surface variables impact some meteorological variables more than others.

Note that the results of this study are potentially dependent on the choice of the models, their spatial resolutions, and the time scales considered in the analysis. Furthermore, the choice of the parameters and of the reference datasets is necessarily subjective given limited computational resources and a variety of different land surface observations. Instead of constraining land model hydrology, future research might consider land surface temperature data or carbon-proxy data and obtain different parameter sensitivities. Another way forward in future studies might be to test spatiotemporal variations of land model parameter (perturbations) instead of applying the same relative changes. Nevertheless, we believe that these caveats do not question the main lessons learned in this study.

Our results moreover have implications on model development strategies. For example, the "tuning" of a coupled system to specific parameterizations of individual model components is somewhat obstructive to exploit its full potential. Improvements in the land surface model may lead to larger temperature and precipitation biases in the coupled system, arising from compensating errors. This highlights the importance of careful testing and calibration of each component prior to the tuning of the entire system. The latter should furthermore take into account multiple skill metrics. Another problem arises from the increasing complexity of many model components and the correspondingly increasing number of parameters. A careful consideration of the role of each parameter is necessary; for instance, if no influence on model performance can be detected even in an extensive analysis, parts of the model code might need to be replaced or reformulated (Rahman et al. 2015). With further increasing model resolutions in future, comprehensive calibration and sensitivity analysis such as introduced in this study might help to cope with the corresponding increase in spatial heterogeneity and local uncertainties.

Acknowledgments. We acknowledge the E-OBS dataset established by the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com, accessed on 24 June 2015) and the data providers in the ECA&D project (http://www.ecad.eu, accessed on 24 June 2015) for sharing precipitation and temperature data, as well as the NASA/GEWEX SRB project (http://gewex-srb.larc.nasa.gov/, accessed on 24 June 2015) and the NASA CERES experiment (http://ceres.larc.nasa.gov/index.php, accessed on 24 June 2015) for sharing radiation data. Furthermore, we acknowledge the international soil moisture network (https://ismn.geo.tuwien.ac.at/, accessed on 24 June 2015) and the SwissSMEX network (http://www.iac.ethz.ch/groups/seneviratne/research/SwissSMEX, accessed on 24 June 2015) for providing soil moisture measurements, Sean Swenson and the NASA MEaSUREs Program for sharing GRACE land data (http://grace.jpl.nasa.gov, accessed on 24 June 2015), the LandFlux-EVAL ET dataset (http://www.iac.ethz.ch/groups/seneviratne/research/LandFlux-EVAL, accessed on 24 June 2015), the European water archive in cooperation with the EU-FP6 project WATCH (http://www.eu-watch.org, accessed on 24 June 2015) for sharing streamflow data, and GPCP data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado (http://www.esrl.noaa.gov/psd/, accessed on 24 June 2015). We thank Frederic Vitart for helpful discussions on the design of the forecasts. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) and European Research Council under Grant Agreements 607085 (EUCLEIA), 603608 (Earth2Observe), 282769 (DROUGHT-R&SPI).

REFERENCES

Adler, R. F., and Coauthors, 2003: The Version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.*, **4**, 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.

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, No. 127, ECMWF, Reading, United Kingdom, 17–22.

——, and Coauthors, 2015: ERA-Interim/Land: A global land water resources data set. *Hydrol. Earth Syst. Sci.*, **19**, 389–407, doi:10.5194/hess-19-389-2015.

Barriopedro, D., E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera, 2011: The hot summer of 2010: Redrawing the temperature record map of Europe. *Science*, **332**, 220–224, doi:10.1126/science.1201224.

Beven, K., 2001: How far can we go in distributed hydrological modelling? *Hydrol. Earth Syst. Sci.*, **5**, 1–12, doi:10.5194/hess-5-1-2001.

——, and A. Binley, 1992: The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Processes*, **6**, 279–298, doi:10.1002/hyp.3360060305.

——, and ——, 2014: GLUE: 20 years on. *Hydrol. Processes*, **28**, 5897–5918, doi:10.1002/hyp.10082.

Bouttier, F., L. Raynaud, O. Nuissier, and B. Ménétrier, 2016: Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Quart. J. Roy. Meteor. Soc.*, doi:10.1002/qj.2622, in press.

Cloke, H. L., F. Pappenberger, and J.-P. Renaud, 2008: Multi-method global sensitivity analysis (MMGSA) for modelling floodplain hydrological processes. *Hydrol. Processes*, **22**, 1660–1674, doi:10.1002/hyp.6734.

——, A. Weisheimer, and F. Pappenberger, 2011: Representing uncertainty in land surface hydrology: Fully coupled simulations with the ECMWF land surface scheme. *Workshop on Representing Model Uncertainty and Error in Numerical Weather and Climate Prediction Models*, Reading, United Kingdom, ECMWF, 109–120. [Available online at http://www.ecmwf.int/sites/default/files/elibrary/2011/8740-representing-uncertainty-land-surface-hydrology-fully-coupled-simulations.pdf.]

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, doi:10.1002/qj.828.

Dobler, C., and F. Pappenberger, 2013: Global sensitivity analyses for a complex hydrological model applied in an Alpine watershed. *Hydrol. Processes*, **27**, 3922–3940, doi:10.1002/hyp.9520.

Fujita, T., D. J. Stensrud, and D. C. Dowell, 2007: Surface data assimilation using an ensemble Kalman filter approach with initial condition and model physics uncertainties. *Mon. Wea. Rev.*, **135**, 1846–1868, doi:10.1175/MWR3391.1.

Gedney, N., C. Huntingford, G. P. Weedon, N. Bellouin, O. Boucher, and P. M. Cox, 2014: Detection of solar dimming and brightening effects on Northern Hemisphere river flow. *Nat. Geosci.*, **7**, 796–800, doi:10.1038/ngeo2263.

Hacker, J. P., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus*, **63**, 625–641, doi:10.1111/j.1600-0870.2010.00497.x.

Harding, R., and Coauthors, 2011: WATCH: Current knowledge of the terrestrial global water cycle. *J. Hydrometeor.*, **12**, 1149–1156, doi:10.1175/JHM-D-11-024.1.

Harrison, K. W., S. V. Kumar, C. D. Peters-Lidard, and J. A. Santanello, 2012: Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques. *Water Resour. Res.*, **48**, W11514, doi:10.1029/2012WR012337.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded dataset of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.*, **113**, D20119, doi:10.1029/2008JD010201.

IPCC, 2013: *Climate Change 2013: The Physical Science Basis*. Cambridge University Press, 1535 pp., doi:10.1017/CBO9781107415324.

Kauffeldt, A., S. Halldin, A. Rodhe, C.-Y. Xu, and I. Westerberg, 2013: Disinformative data in large-scale hydrological modelling. *Hydrol. Earth Syst. Sci.*, **17**, 2845–2857, doi:10.5194/hess-17-2845-2013.

——, ——, F. Pappenberger, F. Wetterhall, C.-Y. Xu, and H. L. Cloke, 2015: Imbalanced land-surface water budgets in a numerical weather prediction system. *Geophys. Res. Lett.*, **42**, 4411–4417, doi:10.1002/2015GL064230.

Koster, R. D., and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, **37**, L02402, doi:10.1029/2009GL041677.

——, and Coauthors, 2011: The second phase of the Global Land–Atmosphere Coupling Experiment: Soil moisture contributions to subseasonal forecast skill. *J. Hydrometeor.*, **12**, 805–822, doi:10.1175/2011JHM1365.1.

Landerer, F. W., and S. C. Swenson, 2012: Accuracy of scaled GRACE terrestrial water storage estimates. *Water Resour. Res.*, **48**, W04531, doi:10.1029/2011WR011453.

Lavaysse, C., M. Carrera, S. Bélair, N. Gagnon, R. Frenette, M. Charron, and M. K. Yau, 2013: Impact of surface parameter uncertainties within the Canadian Regional Ensemble Prediction System. *Mon. Wea. Rev.*, **141**, 1506–1526, doi:10.1175/MWR-D-11-00354.1.

MacLeod, D. A., H. L. Cloke, F. Pappenberger, and A. Weisheimer, 2016: Improved seasonal prediction of the hot summer of 2003 over Europe through better representation of uncertainty in the land surface. *Quart. J. Roy. Meteor. Soc.*, **142**, 79–90, doi:10.1002/qj.2631.

Mueller, B., and Coauthors, 2013: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis. *Hydrol. Earth Syst. Sci.*, **17**, 3707–3720, doi:10.5194/hess-17-3707-2013.

Oleson, K. W., and Coauthors, 2010: Technical description of version 4.0 of the Community Land Model (CLM). NCAR Tech. Note NCAR/TN-478+STR, 257 pp. [Available online at http://www.cesm.ucar.edu/models/cesm1.0/clm/CLM4_Tech_Note.pdf.]

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

——, M. Staudinger, S. I. Seneviratne, J. Seibert, and M. Zappa, 2015: Does model performance improve with complexity? A case study with three hydrological models. *J. Hydrol.*, **523**, 147–159, doi:10.1016/j.jhydrol.2015.01.044.

Pappenberger, F., K. J. Beven, M. Ratto, and P. Matgen, 2008: Multi-method global sensitivity analysis of flood inundation models. *Adv. Water Resour.*, **31**, 1–14, doi:10.1016/j.advwatres.2007.04.009.

Prudhomme, C., F. Doblas-Reyes, O. Bellprat, and E. Dutra, 2016: Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate Dyn.*, doi:10.1007/s00382-015-2879-4, in press.

Rahman, M., M. Sulis, and S. J. Kollet, 2015: The subsurface–land surface–atmosphere connection under convective conditions. *Adv. Water Resour.*, **83**, 240–249, doi:10.1016/j.advwatres.2015.06.003.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, 2008: *Global Sensitivity Analysis: The Primer*. Wiley, 292 pp., doi:10.1002/9780470725184.

Santanello, J. A., Jr., S. V. Kumar, C. D. Peters-Lidard, K. Harrison, and S. Zhou, 2013: Impact of land model calibration on coupled land–atmosphere prediction. *J. Hydrometeor.*, **14**, 1373–1400, doi:10.1175/JHM-D-12-0127.1.

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-Sci. Rev.*, **99**, 125–161, doi:10.1016/j.earscirev.2010.02.004.

——, and Couathors, 2012: Changes in climate extremes and their impacts on the natural physical environment. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, C. B. Field et al., Eds., Cambridge University Press, 109–230.

Sobol, I. M., 1967: On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.*, **7**, 86–112, doi:10.1016/0041-5553(67)90144-9.

Stahl, K., and Couathors, 2010: Streamflow trends in Europe: Evidence from a dataset of near-natural catchments. *Hydrol. Earth Syst. Sci.*, **14**, 2367–2382, doi:10.5194/hess-14-2367-2010.

Swenson, S. C., and J. Wahr, 2006: Post-processing removal of correlated errors in GRACE data. *Geophys. Res. Lett.*, **33**, L08402, doi:10.1029/2005GL025285.

Tennant, W., and S. Beare, 2014: New schemes to perturb sea-surface temperature and soil moisture content in MOGREPS. *Quart. J. Roy. Meteor. Soc.*, **140**, 1150–1160, doi:10.1002/qj.2202.

Touzani, S., and D. Busby, 2013: Smoothing spline analysis of variance approach for global sensitivity analysis of computer codes. *Reliab. Eng. Syst. Saf.*, **112**, 67–81, doi:10.1016/j.ress.2012.11.008.

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**, 1214, doi:10.1029/2002WR001746.

Weedon, G. P., G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo, 2014: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.*, **50**, 7505–7514, doi:10.1002/2014WR015638.

Weisheimer, A., and T. N. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, doi:10.1098/rsif.2013.1162.

Whan, K., J. Zscheischler, R. Orth, M. Shongwe, M. Rahimi, E. Asare, and S. I. Seneviratne, 2015: Impact of soil moisture on extreme maximum temperatures in Europe. *Wea. Climate Extremes*, **9**, 57–67, doi:10.1016/j.wace.2015.05.001.