Effects of Meteorological and Ancillary Data, Temporal Averaging, and Evaluation Methods on Model Performance and Uncertainty in a Land Surface Model

CÉCILE B. MÉNARD, JAAKKO IKONEN, AND KIMMO RAUTIAINEN

Arctic Research Centre, Finnish Meteorological Institute, Helsinki, Finland

MIKA AURELA

Research and Development, Finnish Meteorological Institute, Helsinki, Finland

ALI NADIR ARSLAN AND JOUNI PULLIAINEN

Arctic Research Centre, Finnish Meteorological Institute, Helsinki, Finland

(Manuscript received 21 January 2015, in final form 10 July 2015)

ABSTRACT

A single-model 16-member ensemble is used to investigate how external model factors can affect model performance. Ensemble members are constructed with the land surface model (LSM) Joint UK Land Environment Simulator (JULES), with different choices of meteorological forcing [in situ, NCEP Climate Forecast System Reanalysis (CFSR)/CFSv2, or Water and Global Change (WATCH) Forcing Data ERA-Interim (WFDEI)] and ancillary datasets (in situ or remotely sensed), and with four time step modes. Effects of temporal averaging are investigated by comparing the hourly, daily, monthly, and seasonal ensemble performance against snow depth and water equivalent, soil temperature and moisture, and latent and sensible heat fluxes from one forest site and one clearing in the boreal ecozone of Finnish Lapland. Results show that meteorological data are the largest source of uncertainty; differences in ancillary data have little effect on model results. Although generally informative and representative, aggregated performance metrics fail to identify "right results for the wrong reasons"; to do so, scrutinizing of time series and of interactions between variables is necessary. Temporal averaging over longer intervals improves metrics—with the notable exception of bias, which increases—by reducing the effects of internal data and model variability on model response. Model evaluation during shoulder seasons (fall minus spring) identifies weaknesses in the reanalyses datasets that conventional seasonal performance (winter minus summer) neglects. In view of the importance of snow on the range of results obtained with the same model, let alone identical simulations using different temporal averaging, it is recommended that systematic evaluation, quantification of errors, and uncertainties in snow-covered regions be incorporated in future efforts to standardize evaluation methods of LSMs.

1. Introduction

Over the past two decades, land surface models (LSMs) have evolved from oversimplified schemes, which described the surface boundary conditions for global circulation models (GCMs), to complex models that can be used alone or as part of GCMs to investigate the biogeochemical, hydrological, and energy cycles at the earth's surface (Pitman 2003; Flato et al. 2013). The increasing complexities of LSMs and their broadening applications for climate change studies have warranted a higher scrutiny of the models' internal processes (structure and process parameterizations) and of their associated uncertainties. Numerous model intercomparison projects motivated model improvements by relating model process representation with model structure, focusing on specific aspects of the land surface such as hydrology [e.g., Project for the Intercomparison of Land-Surface Parameterization Schemes (PILPS); Wood et al. 1998; Bowling et al. 2003], snow [e.g., PILPS (Slater et al. 2001), Snow Models Intercomparison Project (SnowMIP; Etchevers et al. 2004), and SnowMIP2 (Essery et al. 2009)], or carbon

Corresponding author address: Cécile Ménard, Finnish Meteorological Institute, Erik Palménin aukio 1, Helsinki 00530, Finland.
E-mail: cecile.menard@ed-alumni.net

DOI: 10.1175/JHM-D-15-0013.1

[e.g., Carbon-Cycle Model Linkage Project (CCMLP; McGuire et al. 2001)]. Other investigators have used single-model platforms, which incorporate and combine multiple process options and parameterizations, in order to eliminate the effects of structural issues when comparing the performance of multiple processes' representations [e.g., Climate High Resolution Model (CHRM; Pomeroy et al. 2007), Framework for Understanding Structural Errors (FUSE; Clark et al. 2008), and the Joint UK Land Environment Simulator (JULES) Investigation Model (JIM; Essery et al. 2013)].

In parallel, the methods employed to evaluate model performance have also come under scrutiny; efforts to establish standardized guidelines for model evaluation have been numerous. For example, Taylor (2001) and Jolliff et al. (2009) have proposed single summary diagrams to represent multiple statistical measures commonly used to quantify model errors. Moriasi et al. (2007), for hydrological models, and Gleckler et al. (2008), for climate models, have recommended the use of specific performance metrics to promote consistency when evaluating different models. Ongoing projects, most notably International Land Model Benchmarking (ILAMB; Luo et al. 2012) and Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER; Best et al. 2015), are focusing on developing internationally accepted frameworks, known as benchmarking frameworks, which aim to facilitate identification of uncertainties in predictions and priorities for future model developments. A more pragmatic consideration, which can considerably affect model performance, is the multitude of user-specified switches, options, and parameters required to determine the optimum model setup. For example, the manual of the community land surface model JULES, version 3.4.1 (JCHMR 2013), lists more than 200 parameters and switches. Although they increase the flexibility and usability of the models for a large community of scientists, they are also potential sources of operational errors as it is becoming increasingly difficult for individual modelers to optimize model states. Luo et al. (2012, p. 3858) states that "it would be unrealistic to expect validation of [the hundreds or thousands of] processes at all spatial and temporal scales independently, even if observations were available." Further to this, we argue that, given the increasing complexity of LSMs, it would also be unrealistic to expect single models to be accurately optimized for simultaneous investigations into hydrological, biogeochemical, and ecological processes, as is expected in climate change studies.

Furthermore, the premise behind many published studies is the implementation or development of process representations that aim to improve model results or, in other words, for results to be closer to observations. While this approach has driven the tremendous progress in LSMs, its shelf life may be limited; some early LSM parameterizations were overly simplistic, and more physically based model developments have led to major improvements (Pitman 2003). For example, the shift in many LSMs from representing snow as part of the upper soil layer for thermal processes to incorporating a physically complex multilayer snowpack dramatically improved the representation of both the energy and water balances at high latitudes (Slater et al. 2001; Pitman 2003; Essery et al. 2013); such major leaps in process representations are likely to be infrequent. Furthermore, the assumption behind this premise is that internal model processes are responsible for "poor" model results; Essery (2013) recently challenged this by showing that much of the uncertainty in the climate–snow albedo feedback in GCMs was due to poor land-cover data, not to choice or complexity of various albedo parameterizations, as was suggested in previous studies.

This paper takes a different approach to evaluating model performance and assessing uncertainty; the aim is to investigate how external model factors, namely, meteorological forcing data, ancillary data, and evaluation methods, affect the "perceived" model performance. In other words, when assessing model performance, we ask two questions:

1) Is the representation of the modeled processes genuinely assessed?
2) Are some of the results artifacts of input data and evaluation methods?

To address these questions, this study focuses on evaluating a limited number of processes in an ensemble of model members that all use the same model structure and parameterizations, one LSM and one single site. The model ensemble is built by varying the meteorological forcing data, the ancillary data, the time step resolution, and the temporal averaging of the results. This approach is deliberately limited in scope in order to facilitate the identification of the source of the differences between members.

2. Description of ensemble study

Uncertainties in modeling studies are expected to rise from four possible sources: 1) input data, 2) interpretation or presentation of model output, 3) model structure and process representation, and 4) parameter values (Renard et al. 2010). The ability of LSMs to investigate the energy, mass, and carbon cycles over large scales means that many investigators rely on reanalysis data to provide the meteorological conditions and on maps, derived from remotely sensed data, to describe the land surface; uncertainty related to input data for regional- or global-scale studies is therefore problematic to quantify. On the other hand, at

TABLE 1. Table of the combinations (crosses) between forcing and ancillary data and time step mode used for each member: FMI, NCEP, and CAP are the three meteorological datasets; IS and CAP are the two sources of ancillary data; and time step modes are represented in the four bottom rows. The asterisks denote the six members that were run hourly, daily, monthly, and seasonally.

	FMI		NCEP		WFDEI	
	IS	CAP	IS	CAP	IS	CAP
1 h	×*	×*				
30-min interpolation	×	×	×	×	×	×
3- to 1-h interpolation	×	×			×*	×*
6- to 1-h interpolation	×	×	×*	×*		

local or "point" scale, LSMs can be forced with automatic weather station data and in situ measurements can inform the parameter values describing the land surface, therefore diminishing, but not altogether removing, uncertainties relating to data source. Interpretation or presentation of model output, although sometimes dictated by the research question itself, rises more from subjective choices and decisions from the modeler, such as how should the model be quantitatively evaluated and which output resolution best represents the natural cycle at which the processes evaluated operate.

To investigate these two sources of uncertainty, as well as discuss the potential contribution of the other two, a 16-member ensemble (Table 1) was built with JULES, each member representing a different combination of the following items:

- Three meteorological forcing datasets: 1) meteorological data measured from automatic weather stations operated by the Finnish Meteorological Institute (FMI); 2) the National Centers for Environmental Prediction Climate Forecast System Reanalysis from 1 June 2007 to 31 December 2010 and Climate Forecast System, version 2, from 1 January 2011 to 15 July 2012 (hereafter, both are referred to collectively as NCEP; Saha et al. 2010, 2014); and 3) the Water and Global Change (WATCH) Forcing Data ERA-Interim (WFDEI; Weedon et al. 2014).
- Two sources for ancillary data: 1) in situ measurements (hereafter IS) of leaf area index (LAI), canopy height, snow-free albedo, and soil texture; and 2) Met Office Central Ancillary Program, version 8.2 (hereafter CAP). Values of the ancillary data are shown in Table 2. All the other vegetation parameters were unmodified from the JULES default name list files.
- Four different time step modes: 30-min time step interpolated from longer time steps (all forcing sources), 1-h when available (FMI only), 1-h time step from interpolation of 6-h (NCEP and FMI), and 3-h (WFDEI and FMI) driving data.

In addition, the effect of temporal averaging on interpretation of the results was queried by investigating the differences between hourly, daily, and monthly output.

3. Site and model description and meteorological and evaluation data

a. Model description

The model used in this study is the JULES land surface model, version 3.4.1, which calculates energy, water, carbon, and momentum exchanges at the earth's surface. The model is the land surface component of the Met Office Unified Model and of the Met Office HadGEM3 GCM; here, it is used in its offline and single-point modes. As JULES proposes a number of user-specific options and parameterizations, the most relevant for this study are described below; the reader is referred to Best et al. (2011) and D. B. Clark et al. (2011) for a full description of the model.

Soil hydraulic properties in this study are taken from Cosby et al. (1984), which relates soil water content, suction, and hydraulic conductivity to soil texture, using the dependencies proposed by Clapp and Hornberger (1978). The number of soil layers and their thickness is user defined and, in this study, follows the default

TABLE 2. List and values of the ancillary data that differ between the CAP and IS members. CAP LAI are from January to December.

	In situ		CAP	
Soil texture	99% sand		78% sand, 9% clay, 13% silt	
b	2.96		11.2	
θ_{crit}	0.08		0.37	
	Clearing	Forest	Clearing	Forest
Albedo	0.18	0.12	0.13	0.13
LAI	0.1	Winter: 0.9, summer: 1.1	0, 0.05, 0.04, 0.17, 0.37, 0.66, 0.87, 0.72, 0.5, 0.7, 0.13, 0	0, 0.16, 0.13, 0.5, 1.09, 1.95, 2.57, 2.15, 1.48, 2.05, 0.37, 0
Canopy height	0.1	14	0.79	21.4

configuration (four layers, from top to bottom, of 0.1, 0.25, 0.65, and 2 m) designed to capture the variation of soil temperature from subdaily to annual time scales (Best et al. 2011). JULES calculates heat fluxes from gradients between temperatures in each layer. The total number of snow layers and the exact thickness of each layer (Δ_k where $k = 1, \ldots,$ maximum number of snow layers) at any one time is determined both by the user and the snow depth S_d. The user-specified thickness of each layer is defined here, from top to bottom, as $0.1, 0.2, 0.2 \leq \Delta_3 \leq S_d$; full details of the snow layers' splitting mechanism are given in Best et al. (2011). Each layer has separate liquid and ice mass, density, and thermal properties. When there is a canopy, snow is either intercepted by the canopy or falls directly on the ground. Within-canopy longwave radiation and sensible heat flux are calculated, but the canopy is treated as opaque for shortwave radiation.

b. Site description

The study site is situated in the principal observation infrastructure of the Finnish Meteorological Institute, the Arctic Research Centre (ARC), located 9 km south of Sodankylä in Tähtelä, Finnish Lapland. The area around Tähtelä, in a 5-km radius, is generally flat (180 ± 5 m above sea level) with vegetation cover typical of boreal environments: 51% forested areas (74% of which are coniferous), 24% wetlands, 14% shrubland and grassland, 5% agricultural areas, 5% urban areas, and 1% lakes (European Environment Agency 2006).

Measurements against which the ensemble was evaluated were collected in two sites situated 60 m from one another and describing two land-cover types: one artificial forest clearing and one forest site (Fig. 1). Automatic S_d, soil temperature T_s, and liquid soil moisture θ measurements are collected at both sites, whereas snow water equivalent (SWE) is only available at the clearing and sensible H and latent (LE) heat fluxes are only available at the forest site. The SWE data are a compilation of automatic measurements from the experimental Astrock Ltd. Gamma Water Instrument (GWI) corrected with snow depth measurements and snow density surveys conducted weekly at the forest clearing. The GWI calculates the snow water equivalent by measuring the rate of gamma ray attenuation in the snowpack at specified spectral bands. The attenuation rate is calibrated a priori using calibration targets, for example, known amount of water or by using reference measurements of SWE. The specifications of the instruments used to collect the evaluation data are given in Table 3.

c. Meteorological forcing data

The model was forced with three different meteorological datasets, all providing incoming shortwave (SW)

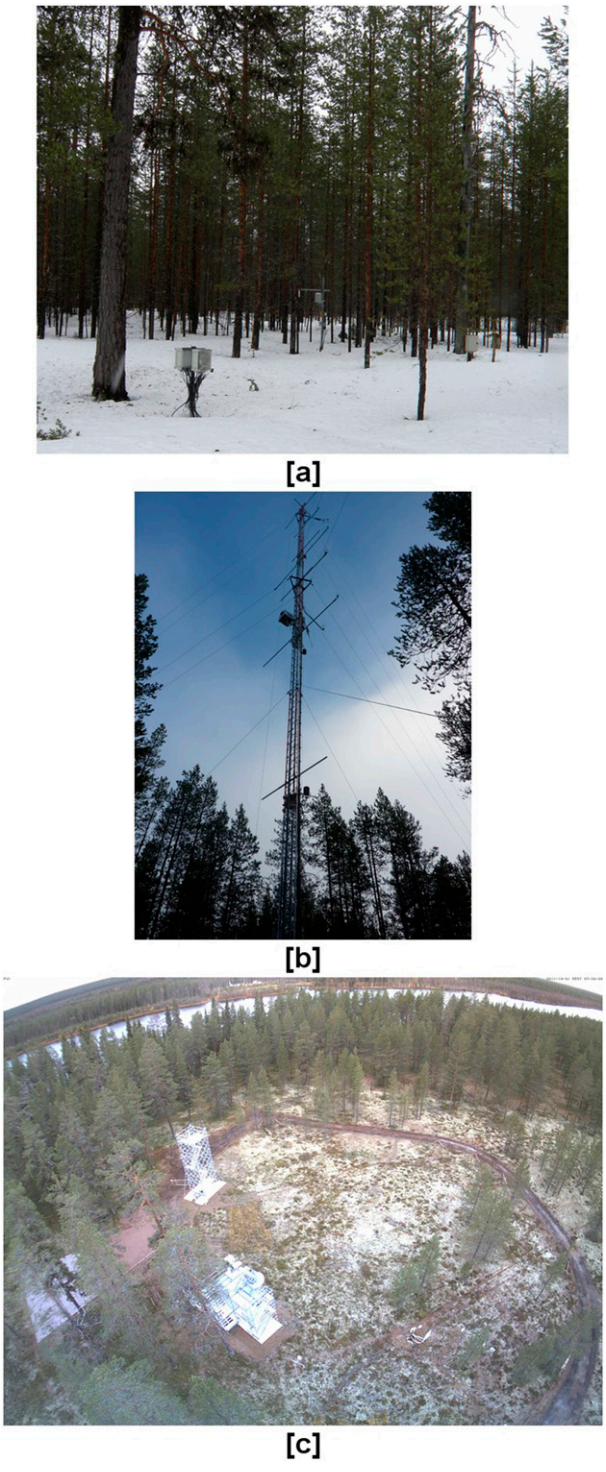

FIG. 1. Photos of (a) soil station and Campbell SR50 sonic ranging sensor, (b) flux tower in the forest, and (c) forest clearing in Tähtelä.

TABLE 3. Specifications of the instruments used to measure the variables against which the ensemble members are evaluated.

Measurement type	Instrument	Accuracy	Number of data gaps (total number of hourly time steps = 34 416)
Snow depth	Campbell SR50 sonic ranging sensor	±1 cm	Clearing: 1704 Forest: 2122
Snow water equivalent	Gamma Water Instrument	±50 mm	20 904
Unfrozen soil moisture	Decagon 5 TE	±3% volumetric water content	2 cm: 2334 10 cm: 2339
Soil temperature	Campbell scientific 109	$<\pm0.2°$ from 0° to 70°C; up to $\pm0.5°$ at −50°C	Clearing: 2342 Forest: 2341
Eddy covariance for latent and sensible heat	LI 700 CO_2/H_2O analyzer	1% in the range of 0–60 mmol mol^{-1}	LE: 18287
	METEK Ultrasonic Anemometer USA-1	0.1 m s^{-1}	H: 18096

and longwave radiation (LW), precipitation P, air temperature T, specific humidity Q_a, wind speed u, and atmospheric pressure (Pa) from 1 July 2007 to 30 June 2012.

All FMI meteorological data used for the clearing simulations are collected at a clearing situated 640 m from the two sites. The station is part of the Global Telecommunication System (GTS), which facilitates the collection, exchange, and distribution of observations within the framework of the World Meteorological Organization. Forest simulations used above-canopy (18 m) T, u, and relative humidity collected at the forest. Acquiring continuous, quality-controlled meteorological driving data is challenging, especially in cold climates, and despite the high level of instrumentation and maintenance at ARC, the data contained numerous gaps ranging from a few hours to a few days. The small (≤3 h) gaps were filled by linear interpolation. Long gaps in longwave radiation measurements before 2007 constrained the start date of the simulations and continued to be problematic, with 2 h of missing data in 2007, 3 h in 2008, 519 h in 2009, 1375 h in 2010, 1663 h in 2011, and 3 h in 2012. The gaps exceeding 3 h were filled following Wunderlich (1972) to diagnose incoming longwave radiation based on fractional cloud cover, air temperature, and atmospheric vapor pressure; the emissivity of clouds and fractional cloud cover computations are based on algorithms described in Deardorff (1978).

For comparison with in situ meteorological data, the corresponding grid box in the NCEP and WFDEI datasets was selected. Both products are 0.5° × 0.5° resolution (67°–67.5°N and 26.5°–27°W), covering 55.6 km of latitude and 21.7 km of longitude. No attempt to downscale the reanalysis data was performed, as downscaling usually involves using measured data; this study aims to compare model results between reanalysis and measured data instead. The NCEP 6-hourly product was chosen because it is the only combination of the NCEP-CFS series that provided all the forcing data at the same resolution from 2007 to 2012. The WFDEI 3-hourly dataset is based on the European Centre for Medium-Range Weather Forecasts interim reanalysis data (Dee et al. 2011) and was compiled to provide easily accessible forcing data for land surface models (Weedon et al. 2014). The data extend from 1979 to 2012, which is why this study assesses model performance until 2012.

Table 3 shows a comparison, rather than an evaluation, because the FMI data are representative of a smaller area than the grid box of the reanalyses products, between the NCEP and WFDEI datasets and the FMI station. The statistics show that the two reanalysis datasets are good proxies of local meteorological conditions. ERA-Interim assimilates FMI measurements of air temperature and relative humidity via the GTS, which explains the small WFDEI root-mean-square error (RMSE), small bias, and high correlation coefficient R. In general, both reanalysis products show high correlation and low bias; RMSE is generally low but highest with NCEP. The notable exceptions are the large bias and RMSE in precipitation, and to a lesser extent wind speed, with the NCEP data. As documented by Dee et al. (2011) and Weedon et al. (2014), these two variables are also generally the most problematic in ERA-Interim; despite a low bias, R for the two variables are the lowest and the ratio of RMSE to standard deviation of observations σ_O, sometimes used to quantify model performance (Moriasi et al. 2007), is the highest.

d. Ancillary data

Site-specific measurements of LAI, snow-free albedo, and soil texture properties provided the values for the ancillary data used in the IS simulations (Table 2). Winter LAI for the forest site was calculated from

ground-based measurements summarized in Reid et al. (2013), which agreed with values in Manninen et al. (2012). Summer LAI was defined as being 20% higher than winter LAI following Manninen et al. (2012). In the absence of year-round measurements and given the small difference between summer and winter LAI, the summer value was used from June to August and the winter value was used for the rest of the year. No LAI measurements are available for the forest clearing, which is characterized by sparse short grass; a nominal value of 0.1 was used. Snow-free albedo was measured with albedometers installed at both sites in summer 2012. A soil survey conducted in 2013 determined the mineral soil at both sites to be composed of approximately 99% sand; organic matter was found to be <5% up to 20-cm depth and <2% in deeper layers and was therefore ignored.

The Central Ancillary Program was developed by the Met Office (United Kingdom) and creates ancillary data using globally available datasets. No official documentation exists, although the CAP data are used in the Met Office Unified Model, of which JULES is the land surface model. CAP LAI is derived from monthly means between 2005 and 2009 of the Moderate Resolution Imaging Spectroradiometer (MODIS) global 4-km LAI product. Soil texture is derived from the Harmonized World Soil Database (HWSD) supplemented with other regional soils datasets at 1-km resolution. Albedo for each plant functional type in JULES was derived by Houldcroft et al. (2009) from the MODIS white sky albedo product.

e. Evaluation methods

The ensemble members are evaluated using metrics that summarize different aspects of model performance: bias evaluates the accuracy; R evaluates the agreement in temporal patterns; and the RMSE and standard deviation σ_M evaluate the amplitude in the variation from the observations and from the mean, respectively. The unbiased RMSE (URMSE), R, and σ_M are plotted in Taylor diagrams, following Taylor (2001), who used the relationship between URMSE, R, σ_M, and the law of cosines to construct a polar summary diagram. Each variable is normalized by the standard deviation of the corresponding observations, thus allowing multiple variables with different dimensions to be shown on the same plot; errors depend on the magnitude of the variations in the observations and thus will be larger for larger σ_O. In Figs. 2–7, the reference point, representing observations, is situated at URMSE = 0, R = 1, and normalized σ_M = 1. URMSE is represented as the radial distance from the reference point, normalized σ_M as the radial distance from the origin, and R as the azimuthal angular position. Using the relationship $\mathrm{RMSE}^2 = \mathrm{URMSE}^2 + \mathrm{bias}^2$,

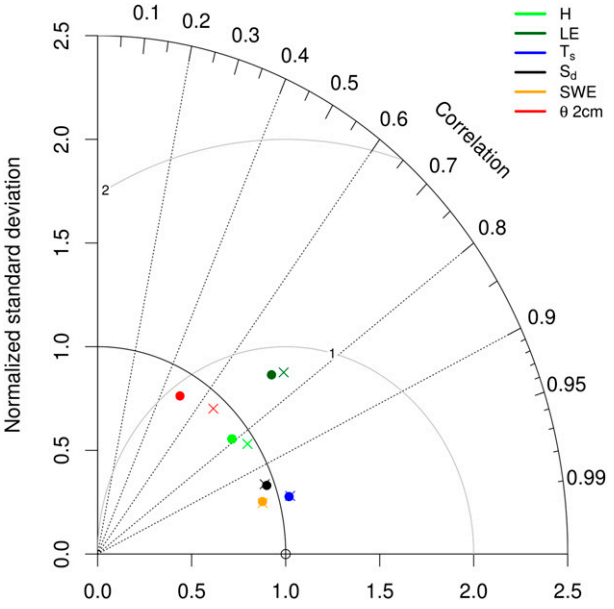

FIG. 2. Differences in pattern statistics between members forced with 1-h interval meteorological data (dots) and with 6-h-interpolated to 1-h data (crosses). All results use FMI meteorological data.

described by Taylor (2001), stacked bar plots show the relative contribution of the bias and the URMSE to the RMSE. Time series of the ensemble for each variable evaluated against measurements are also shown (Figs. 3–6).

Results of the members marked with an asterisk in Table 1 were averaged over three different time scales: hourly, daily, and monthly. Figures 3c, 4c, 5c, and 6c show biases for all members, and Fig. 7 shows differences in performance metrics for variables affected by differences in temporal averaging. For simplicity, Fig. 7 only shows IS members because there was no significant difference between IS and CAP members.

Model uncertainty that arises from the different components constituting each member is quantified by adapting the method proposed by Déqué et al. (2007), which uses the variances between individual members against measurements.

Defining X_{ijk} as one of the model output variables where index i = 1–3 for the number of forcing meteorological datasets (FMI, WFDEI, and NCEP), j = 1–3 for the number of temporal output resolution (hourly, daily, and monthly), and k = 1–2 for the number of ancillary datasets (IS and CAP), the variance V in X_{ijk} can be decomposed as

$$V(X_{ijk}) = X_{\mathrm{obs}} + F_i + O_j + A_k + (FO)_{ij} + (FA)_{ik} + (OA)_{jk} + (FOA)_{ijk}, \quad (1)$$

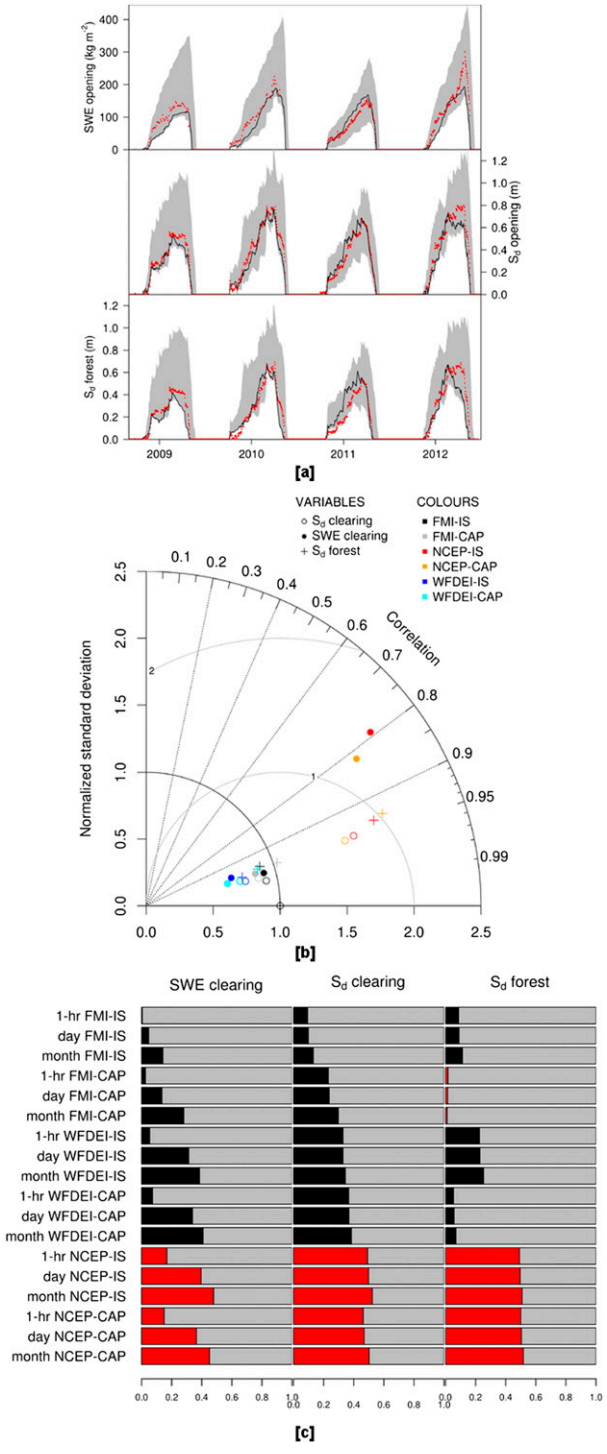

FIG. 3. Measures of model performance (a) range (gray band) of observed (red) and modeled (black) snow depth: (from top to bottom) SWE opening, S_d opening, and S_d forest. (b) SWE Taylor diagram of model ensemble performance. Members are differentiated by their color (meteorological forcing + ancillary data) and symbol (variable). For example, the red circle represents the performance of the modeled snow depth at the clearing for ensemble NCEP-IS. (c) Stacked bar plot of the relationship between RMSE, URMSE, and bias such that $\text{RMSE}^2 = 1 = \text{URMSE}^2 + \text{bias}^2$ for (from left to right) SWE clearing, S_d clearing, and S_d forest. Black and red bars indicate the ratio of bias^2 to RMSE^2 when the bias is negative and positive, respectively; gray bars indicate the ratio of URMSE^2 to RMSE^2.

where X_{obs} is the observation; F is the individual part of the variance due to the origin of the forcing data i; and O and A are as previously given, but for output resolution j and ancillary data k, respectively. The other terms represent the parts of the variance due to the interaction effects of respective terms; equations for each term are presented in Table 5. Uncertainty from each contributing factor in this study is presented in Table 6 and referred to in the following sections as percentage contributions of the total variance.

4. Results

Model steady state was obtained by spinning up the first year of data, which was not used in the evaluation study, multiple times until equilibrium was reached.

a. Effect of time step interpolation on model results

JULES is known to be numerically unstable for time steps longer than 60 min and converts meteorological datasets with long intervals to a user-defined time step length (here 30 min and 1 h) by linearly interpolating between the two given times while preserving the period means of the driving data file (Sheng and Zwiers 1998). Although this step is not always acknowledged in the literature, it is compulsory for all JULES simulations that use reanalysis data with time scales >1 h.

The Sheng and Zwiers (1998) scheme was initially developed to obtain daily values from monthly means; by proposing an approach that preserved monthly mean, they prevented artificial steps at the joining of calendar months (Vincent et al. 2002). No literature on the effect of the interpolation procedure for subdaily time steps exists; a complete evaluation of the scheme is also beyond the scope of this study. Nevertheless, in order to assess the potential effects of time interpolation on WFDEI and NCEP results, hourly FMI meteorological measurements were averaged over 3 and 6 h to construct one 3-h and one 6-h interval forcing dataset. The FMI, WFDEI, and NCEP datasets were also interpolated to 30-min time steps to assess whether shorter time steps provided a more accurate representation of the processes evaluated. For simplicity, results of the original

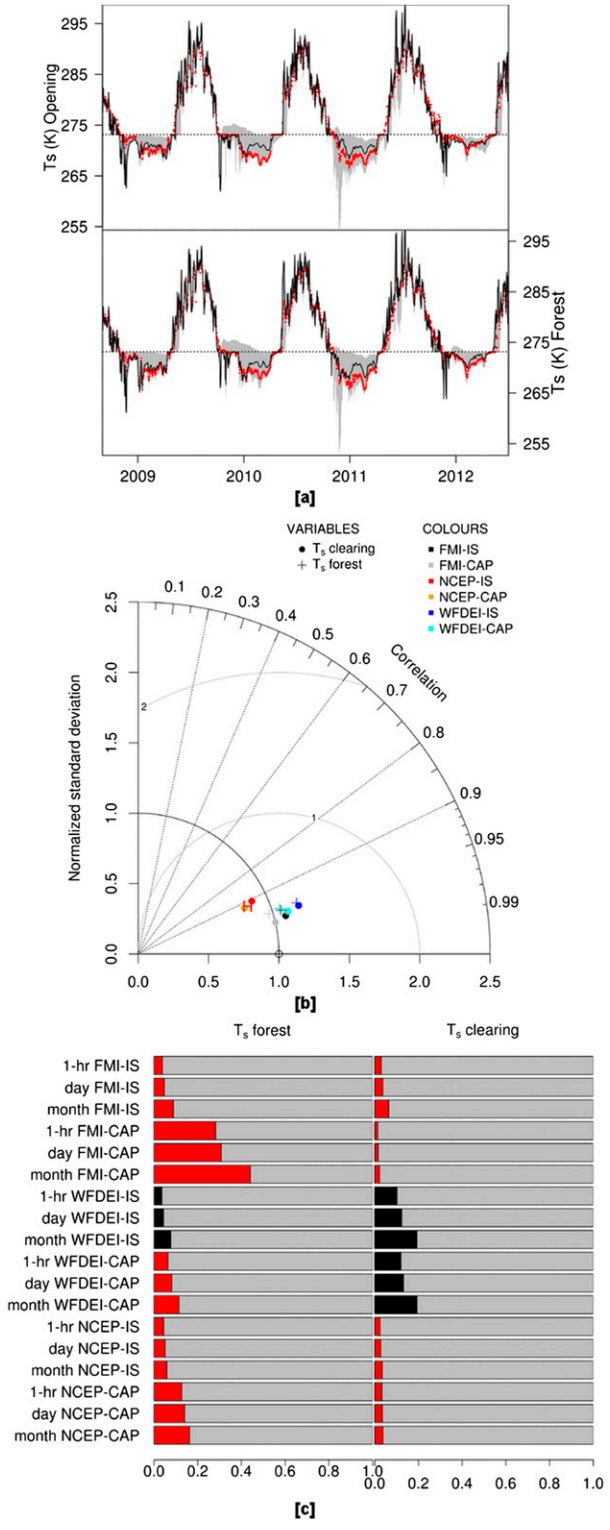

FIG. 4. As in Fig. 3, but for soil temperature.

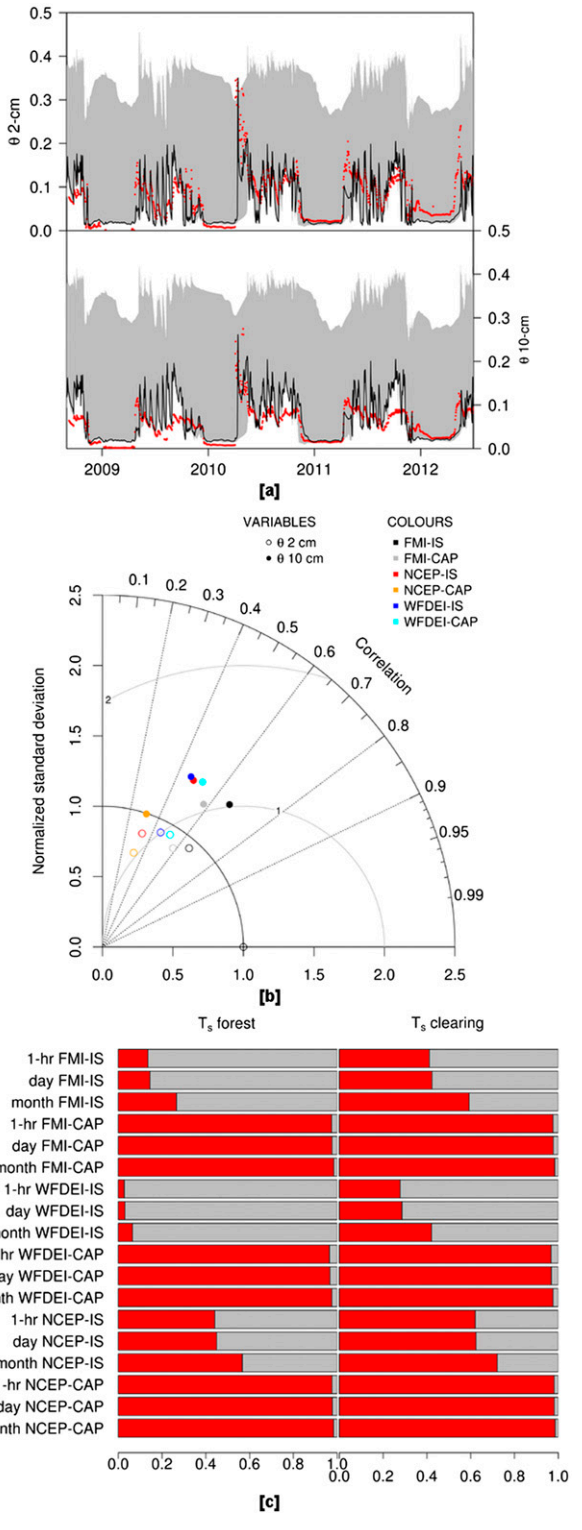

FIG. 5. As in Fig. 3, but for unfrozen soil moisture at 2- and 10-cm depths (forest clearing site only).

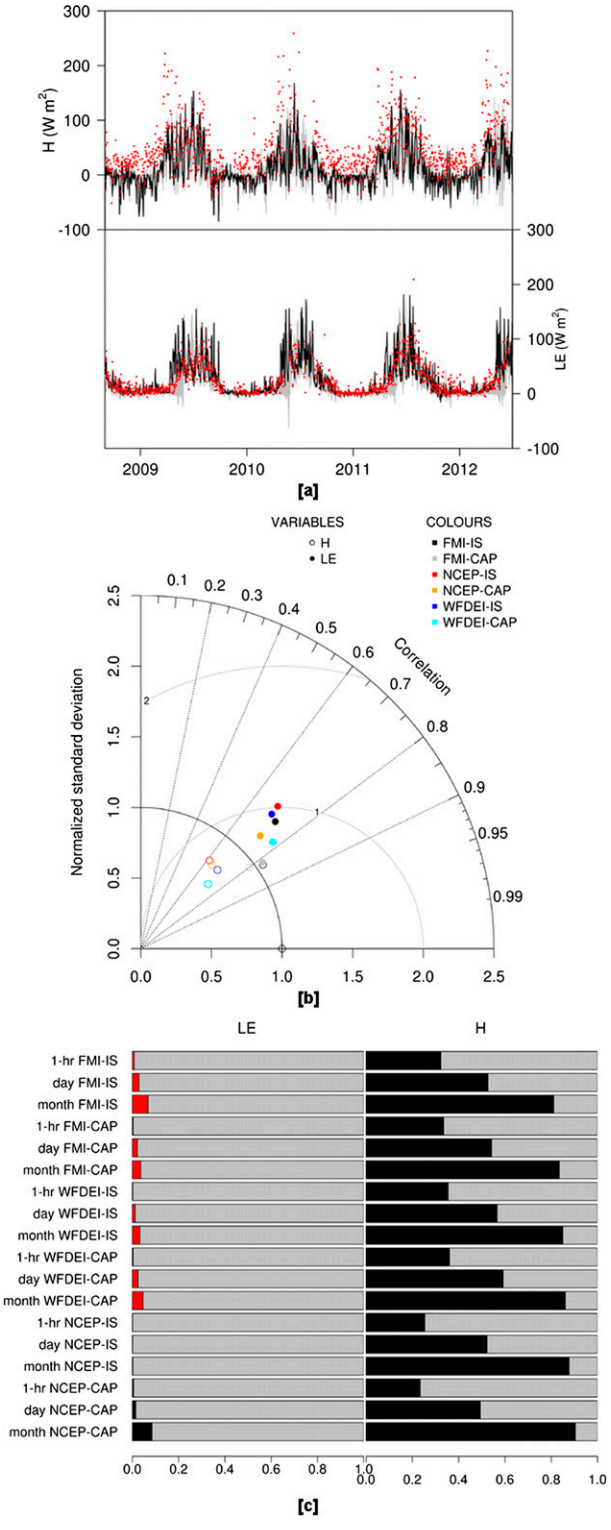

FIG. 6. As in Fig. 3, but for sensible and latent heat fluxes (forest site only).

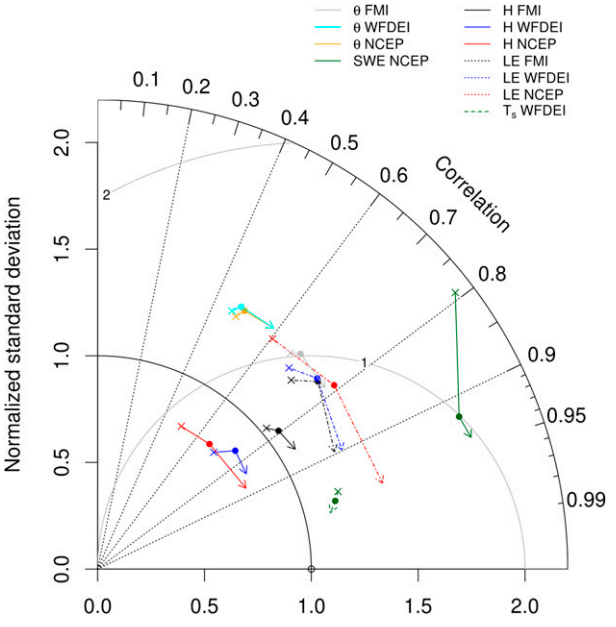

FIG. 7. Changes in performance statistics with increasing times scales. The arrow symbols show: the base (crosses) hourly output, the middle (circle) daily output, and the head monthly output. The LE and H are from the forest; SWE and θ at 10 cm are from the clearing.

1-h FMI dataset and the 6-h dataset, then interpolated to 1-h by JULES, only are compared in Fig. 2.

There were no noticeable differences between the 30-min and 1-h members; for simplicity, 1-h members-only are shown in Fig. 2. The time interval of the driving data has no effect on T_s, S_d, and SWE and makes little difference to H and LE, if only a small decrease in standard deviation. Soil moisture is the variable most affected, with a decrease in R of 0.15 (0.65–0.5) suggesting that the interpolation scheme may be more appropriate to calculate surface heat transfers than soil hydraulics.

The small scale of the differences between members in Fig. 2 suggests that comparing model results from members with driving data of different time intervals should not affect the interpretation of the results. As a consequence, all the results discussed in the next sections use 1-h time steps only (members marked with an asterisk in Table 1).

b. *Effects of forcing and ancillary data on model results*

1) SNOW DEPTH AND WATER EQUIVALENT

The performance of the ensemble in simulating snow depth and snow water equivalent is shown in Fig. 3. The FMI-IS member reproduces timing and patterns of

TABLE 4. Comparison between the NCEP and WFDEI datasets and the FMI meteorological data. The standard deviation and mean of the FMI meteorological variables are given as an indication of the distribution of the measured values.

		NCEP			WFDEI		
	σ_O (mean)	RMSE	Bias	R	RMSE	Bias	R
Incoming SW ($W\,m^{-2}$)	154 (94)	79	−0.13	0.87	52	−0.05	0.88
Incoming LW ($W\,m^{-2}$)	52 (283)	30	0.02	0.84	25	−0.01	0.79
Specific humidity ($kg\,kg^{-1}$)	2.5×10^{-3} (3.9×10^{-3})	6.9×10^{-4}	0.08	0.97	7.5×10^{-4}	0.04	0.92
Precipitation ($kg\,m^{-2}\,day^{-1}$)	3.36 (0.76)	3.06	0.64	0.66	2.56	0.01	0.68
Wind speed ($m\,s^{-1}$)	1.4 (2.39)	1.07	0.22	0.8	0.85	−0.02	0.64
Air temperature (°C)	11.42 (0.26)	3.45	0	0.96	3	0	0.92
Pressure (Pa)	1206 (98 809)	631	−0.01	1	621	−0.01	1

accumulation and melt very well; maximum difference in the first snow-free day each year between measurements and FMI-IS is 3 days. Correlation for all members is high, with all members exceeding 0.78 and all but three exceeding 0.94. All NCEP members have larger URMSE, standard deviation, and bias than the other members, showing that they overestimate both the mean and amplitude of changes in S_d and SWE. This is consistent with errors in precipitation data in Table 4; the amount of snow on the ground is overestimated (bias) and, by extension, so is the snow accumulation between each snowfall (URMSE and σ_M). Figure 2c shows little difference in the contribution of the bias to the RMSE (hereafter ''bias'' for short) between CAP and IS members at the clearing. In the forest, FMI and WFDEI snow is deeper with CAP members because of the lower LAI, which reduces both sublimation rates and canopy temperature. In JULES, maximum canopy snow load equals 4.4 LAI (Essery et al. 2003) and, with a higher LAI, IS members intercept and sublimate more snow than CAP members. In addition, the heat capacity of the canopy depends on leaf and wood biomass, both of which are related to LAI. Half of the difference in snow depth (and SWE, not shown) between CAP and IS members occurs at the beginning and end of the snow season, when air temperature is close to melting point, but canopy temperature is higher. A higher LAI generates a hotter canopy and, by extension, promotes turbulent exchanges and longwave radiation and, therefore, melt beneath the canopy. The effect of LAI is also seen in Table 5, where all sources including ancillary data are higher in the forest than in the clearing.

The negative bias in WFDEI S_d and SWE does not correspond to the very small bias in WFDEI precipitation data (Table 4). In fact, WFDEI underestimates snowfall in January–February 2010 and for most of the 2010/11 season (not shown). At the end of the 2008/09 snow season, the total WFDEI snowfall is 9% lower than FMI; by the end of 2010/11, the total difference has doubled. Table 3 also shows that, although there is a very low bias in precipitation, the normalized RMSE (RMSE/σ_M) is relatively high (0.76) and R is relatively low compared to other variables, corroborating that the underestimated snowfall is compensated for by overestimated summer rainfall.

2) SOIL TEMPERATURE

Soil temperatures are generally close to measurements; all simulations have very high R (>0.9), small URMSE (<0.5), and σ_M within 20% of σ_O (Fig. 4). The deepest snow in Fig. 3a and highest T_s in Fig. 4a are from

TABLE 5. Equations detailing each term in Eq. (1). Variable X is the modeled variable (e.g., snow depth), i = 1–3 according to the number of meteorological datasets (F), j = 1–3 according to the number of temporal output resolutions (O), and k = 1–2 according to the number of ancillary datasets (A). For example, F is the contribution of the forcing data to the variance in the ensemble and FO is the contribution of the combined interaction between the forcing data and temporal averaging of the results. The full part of the variance due to forcing data would be described as $X_{ijk}(F) = F_i + (FO)_{ij} + (FA)_{ik} + (FAO)_{ijk}$, but the total variance is not the sum of $X_{ijk}(F) + X_{ijk}(A) + X_{ijk}(O)$ because some of the terms are repeated [e.g., $(FO)_{ij}$ is in both $X_{ijk}(F)$ and $X_{ijk}(O)$].

Terms in Eq. (1)	Variance equation
F	$\sum_{i=1}^{3} \frac{(X_i - X_{\text{obs}})^2}{3}$
O	$\sum_{j=1}^{3} \frac{(X_j - X_{\text{obs}})^2}{3}$
A	$\sum_{k=1}^{2} \frac{(X_k - X_{\text{obs}})^2}{2}$
FO	$\sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(X_{ij} - X_i - X_j + X_{\text{obs}})^2}{9}$
FA	$\sum_{i=1}^{3}\sum_{k=1}^{2} \frac{(X_{ik} - X_i - X_k + X_{\text{obs}})^2}{6}$
OA	$\sum_{j=1}^{3}\sum_{k=1}^{2} \frac{(X_{jk} - X_j - X_k + X_{\text{obs}})^2}{6}$
FOA	$\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{2} \frac{(X_{ijk} - X_{ij} - X_{ik} - X_{jk} + X_i + X_j + X_k - X_{\text{obs}})^2}{18}$

TABLE 6. Relative contribution to uncertainties, expressed as percentages of the total variance, of the forcing data, temporal averaging, ancillary data, and their combined effects on the evaluated variables. The F denotes forest and C denotes clearing. Boldface values denote the largest contribution to uncertainty for each variable.

	Forcing data	Temporal averaging	Ancillary data	Forcing + output	Forcing + ancillary	Output + ancillary	All
S_d (C)	**69**	6	5	5	5	5	5
S_d (F)	**42**	10	10	9	10	9	10
SWE	11	**59**	2	17	2	2	7
T_s (C)	**94**	4	0	0	1	0	1
T_s (F)	16	14	16	11	12	11	**20**
θ 2 cm	13	13	**18**	13	13	13	17
θ 10 cm	12	12	**21**	12	12	12	19
H	14.3	14.3	14.3	14.3	14.3	14.3	14.3
LE	**59**	6	9	6	8	5	6
Average	**37**	15	11	10	8	8	11

the NCEP members, showing that snow insulation dominates winter soil temperatures. Of all the evaluated variables, T_s at the clearing is the most sensitive to forcing data; at the forest it is the only variable for which the combination of the three different sources weighs more than the individual sources. WFDEI and FMI members have higher T_s with CAP data, which affect snow depth and, by extension, insulation, but also soil properties; NCEP members are less sensitive to changes in ancillary data because the large bias in snow amount dominates the soil thermal regime. NCEP members also underestimate σ_M because, as the deep snow insulates the soil, temperatures are less sensitive to changes in meteorological conditions and remain relatively high year-round; this is discussed further in section 4d.

Figure 4a shows that, at shallow snow depths at the beginning of the snow season, soil temperatures are underestimated and can fluctuate rapidly between 262 and 273.15 K. This fluctuation is a result of a numerical artifact; to avoid numerical instabilities when $0 < S_d < \Delta_1$, the top soil layer in JULES becomes a soil–snow composite that thermally functions as a single layer. The temperature of this layer is taken at a fixed depth below the surface whether snow is present or not. As a consequence, soil temperature, at 10 cm or less, represents either the soil or the snow temperature, depending on the exact snow depth.

3) UNFROZEN SOIL MOISTURE

In Fig. 5b, with the exception of NCEP-CAP, all σ_M at 2-cm depth are within 0.15 of σ_O; at 10 cm, all but NCEP-CAP members overestimate σ_M by a minimum of 0.24, showing that the model underestimates dampening of signal with depth. At specific depths and with the exception of NCEP-CAP at 2 cm, the scatter in URMSE and σ_M is relatively small. The value of R doubles between NCEP and FMI members; this difference may be partly due to the larger time scales in the driving data, as Fig. 2 showed that θ was the variable the most affected by the time interpolation scheme. The two θ are the only variables in Table 6 whose main source of uncertainty is the ancillary data. The large positive bias with CAP members in Figs. 5a and 5c is due to the Clapp and Hornberger (1978) b parameter to which JULES is particularly sensitive. The variable b is used to calculate soil water suction and the hydraulic conductivity at saturation; higher b values increase the minimum volume of water present in the layer such that, with $b = 11.2$ with CAP members, θ does not fall below 0.3 of the saturated fraction against 0.02 with IS members when $b = 2.96$. The positive bias with NCEP members occurs because snow amount and soil temperatures are overestimated and, therefore, a larger fraction of soil moisture remains unfrozen during winter.

4) TURBULENT FLUXES

In Fig. 6b, the standard deviation is underestimated in H but overestimated in LE. The difference in LE between CAP and IS members occurs mostly during the snowmelt season, when the IS members overestimate latent heat fluxes but CAP members do not. The perceived increased performance occurs because soil conductance is inversely proportional to the critical point at which soil moisture stress starts to restrict transpiration and that depends on soil texture (Table 2); although CAP members diagnose a higher θ than IS members, with a lower critical point less moisture is available for evaporation. Unlike in Blyth et al. (2010), LAI does not affect LE.

All members have a negative H bias. Equal weight between sources in Table 6 suggests that the main source of uncertainty may rest elsewhere, for example, in the data or process representation. The latter would be consistent with Best et al. (2015), who suggested that the conceptual view of energy partitioning in LSMs is essentially flawed and the equation representing surface fluxes is incorrect. Instrument errors may also be responsible for model uncertainties: the anemometer used

to calculate sensible heat fluxes (Table 3) was replaced by another ultrasonic anemometer (METEK USA-1) in November 2013. A comparison of H between the two instruments shows lower sensible heat fluxes in 2014 than in previous years. Average H between 2008 and 2013 ranged between 40 and 48 W m^{-2} against 30 W m^{-2} in 2014; the average median was between 14 and 28 W m^{-2} before the replacement against -2 W m^{-2} in 2014. The differences are larger in winter but, as of now, the exact cause of the possible overestimation of the measured sensible heat fluxes has not been identified. A longer time series with the new instrument may help elucidate this bias in the data and/or model.

c. *Effect of temporal averaging of results*

Strictly speaking, comparing model results and observations at hourly scale is only valid for FMI members, as the NCEP and WFDEI datasets are only available at intervals greater than 1 h. However, even for daily and monthly output, JULES does not distinguish between actual and interpolated data such that model results produced at any time interval always used results from interpolated data. As a consequence, when evaluating JULES over any time scale when driven by reanalysis datasets, the interpolation procedure is always implicitly evaluated.

By reducing the effects of data and model variability on model response, quantitative errors are generally smallest at monthly time scales. Figure 7 shows a systematic decrease in URMSE and increase in R, with the largest differences occurring for NCEP H, LE, and SWE. However, the improvements in URMSE and correlations occur at the expense of bias (Figs. 3c–6c), which consistently increases with increasing output resolution. This corroborates Decker et al. (2012), who found that the dominant error in reanalysis datasets (including ERA-Interim and NCEP CFSR) was correlation at the time step resolution but bias at monthly resolution. They stressed that future developments in the production of reanalysis data should not only focus on the well-known monthly biases in precipitation and temperature (although the latter was not found in this study) but also on solving correlation errors at shorter time scales.

Given that forcing data are the largest source of uncertainties in S_d in Table 6, it is, at first, surprising that uncertainty in SWE is mostly attributed to output resolution. These differences are, actually, artifacts of the evaluation method. The length of the snow season is longer with the NCEP members; in order to treat all members equally, year-round SWE, that is, including when SWE = 0, was used to calculate errors. Table 3 shows that 61% of hourly SWE measurements are missing. In fact, at hourly resolution, the sum of missing data and SWE = 0 equals 97% of the data, with average SWE = 12 kg m^{-2}. By averaging measurements for daily and monthly resolution, the proportional number of missing data decreases and the number of time steps with SWE > 0 increases to 29% and 58%, respectively; the number of snow-free time steps weighs less on the distribution of SWE, which averages 47 and 54 kg m^{-2} daily and monthly, respectively.

d. *Is "seasonal" performance represented?*

Seasonal performance of models is conventionally assessed by examining the difference between means for December–February (DJF) and June–August (JJA; Flato et al. 2013). However, these 3-month calendar-based seasons do not reflect the dynamics of the long winters in Finnish Lapland and in other high-latitude locations. Here, the calendar-based seasonal performance is compared to "local" seasons, all defined according to the yearly snow cycle: 1) winter (the snow cover season; November–March), 2) summer (the snow-free season; May–September), 3) spring (the melt season; April), and 4) autumn (accumulation season; October). Spring and autumn are also often called the "shoulder" seasons. This approach is consistent with one of the hypothesis testing approaches proposed by M. P. Clark et al. (2011), who argue for model evaluation against significant hydrological behavior rather than mere matching of model against observations.

Figure 8 shows that considering local seasons, which have more warm months than calendar-based seasons, decreases the seasonal amplitude in all but θ. Differences between model results and measurements are of similar magnitude, independently of the season definition, for all variables but the latent heat fluxes. There is a larger spread in modeled seasonal amplitude in LE with the calendar-based seasons (4–20 W m^{-2} against 10–16 W m^{-2}), showing that members with smaller errors compared to local seasons (FMI-IS and WFDEI-IS) are better at simulating processes during temperature extremes rather than during the warmer snow-covered months.

Differences in observations between calendar and local seasons are reduced during the shoulder seasons (Fig. 9). There is no bias in sensible heat fluxes with the FMI and NCEP members during the shoulder months, further strengthening the hypothesis of a malfunctioning instrument in winter. Unlike in Fig. 8, the WFDEI members underestimate differences in H in Fig. 9; springtime is the time of the year when the land surface is most heterogeneous because of patchy snow cover and, by extension, the time of the year when scale mismatch may affect results most. However, separating Table 2 into seasons (not shown) does not reveal a specific season during which the WFDEI data are more prone to errors. The effect on soil temperatures of the

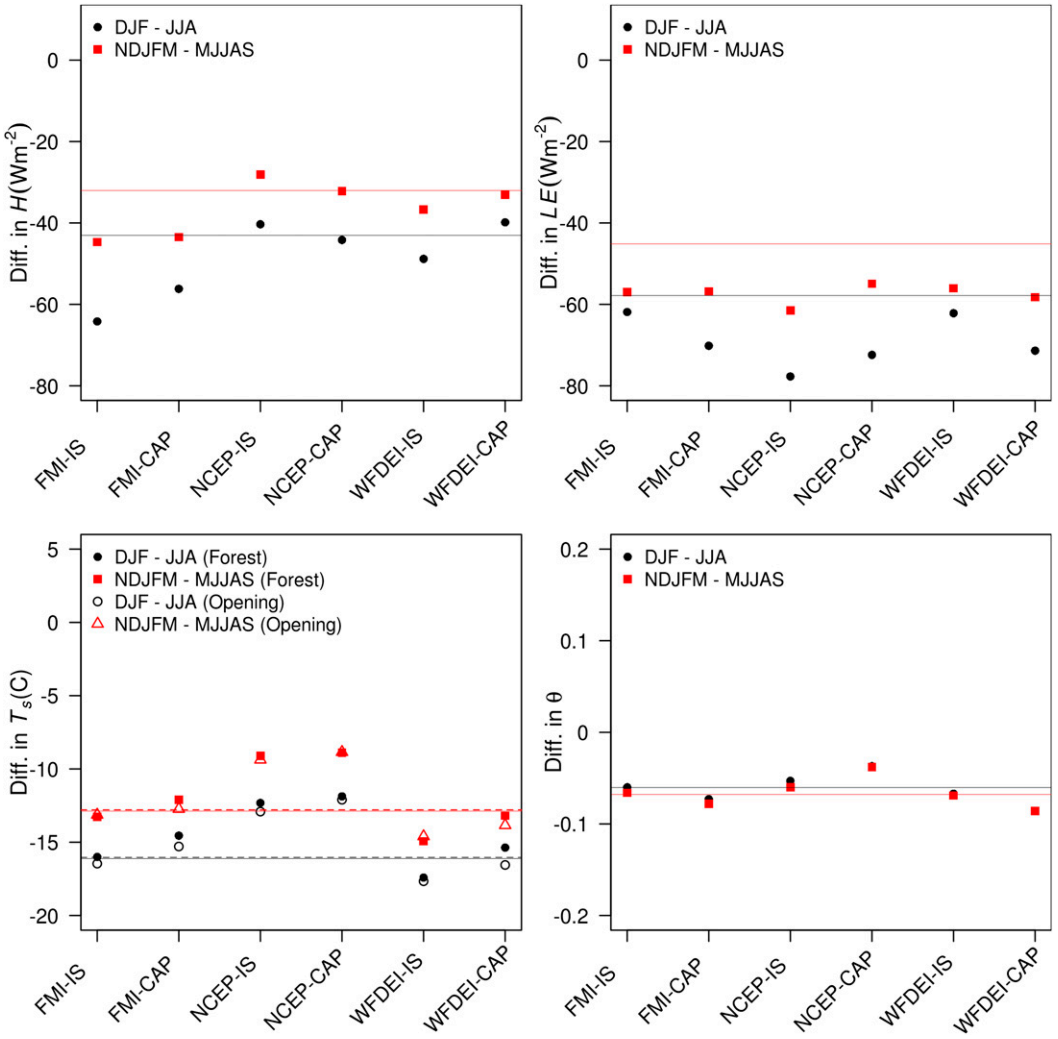

FIG. 8. Modeled (symbols) vs observed (horizontal lines) seasonal (winter minus summer) differences in (top) (left) sensible and (right) latent heat fluxes; and (bottom) (left) soil temperature, and (right) unfrozen soil moisture. Letters in the legend refer to months of the year. Measured and modeled calendar-based seasons (DJF minus JJA) are shown in black; local seasons (NDJFM minus MJJAS) are in red. Solid lines show measured differences in the forest; dashed lines in T_s are for the forest opening.

overestimated NCEP precipitation is even clearer in Fig. 8 than in Fig. 4. First, seasonal difference shows that T_s is underestimated by up to 5.4°C because deep snow insulation favors warm winter soil temperatures. Second, during the shoulder seasons in Fig. 9, NCEP members appear to perform consistently well; the mean differences in LE are the closest to measurements in the ensemble. In fact, these are "good results for the wrong reasons." In April, NCEP members have deeper snow than other members and snow cover extends, on average, 2 weeks with CAP data and 18 days with IS data beyond observed melt; this offsets the positive bias in modeled moisture fluxes during melt because the top soil layer is wet and moisture is available for evaporation, as happens when modeled snow is close to observations. This process explains the dominance of forcing data on uncertainties in LE in Table 6.

5. Discussion and conclusions

This study presented a 16-member ensemble, built with different forcing and ancillary data to investigate errors and uncertainties in a single model at a single site over different temporal scales. The choice of the model itself was dictated by the familiarity of the authors with JULES, but the issues discussed here are not model-specific; interpretation of model performance concerns the land surface model community as a whole. Furthermore,

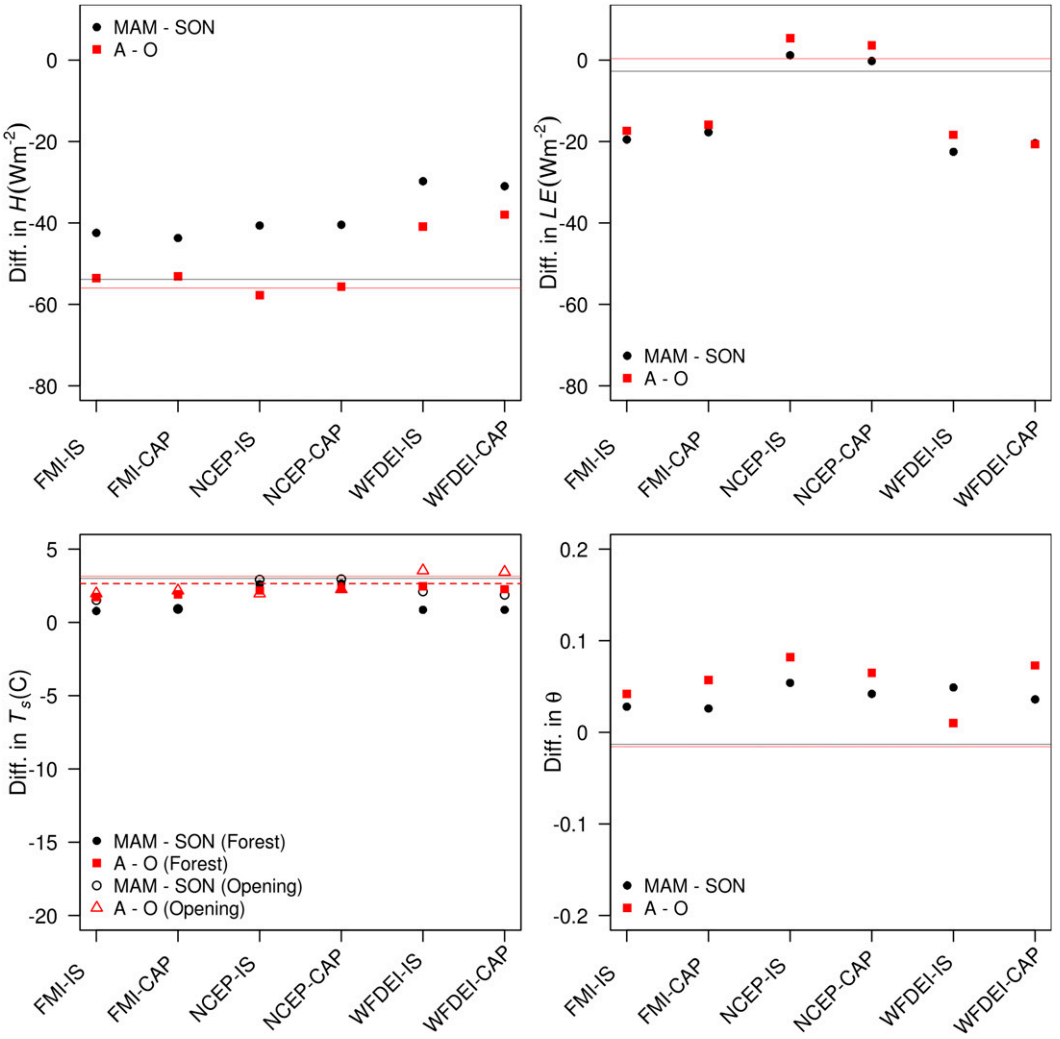

FIG. 9. As in Fig. 8, but for measured and modeled calendar-based shoulder seasons (March–May minus September–November) shown in black; local shoulder seasons (April minus October) in red.

while, in principle, using a single site for this investigation may appear limited in scope, the single-site approach allowed identification of errors in the interpretation of model performance that spatially broader studies would likely have missed.

The design of this study was determined by the necessity to answer the first question raised in the introduction positively in order to address the second question. The short answer to question 1 is "yes," inasmuch as "good" performance is defined here as an absence of significant biases and a good correspondence between the modeled amplitude and seasonality and measurements at the studied site when provided with measured meteorological and ancillary data; it is worth noting that Best et al. (2015) found that models that performed "well" when evaluated against observations, as is the case here, can still be found to perform poorly when a priori benchmark levels for particular metrics are defined. Model performance when using reanalysis data is generally poorer, but differences in ancillary data had little effect on model results. The poorer performance of NCEP and WFDEI members compared to the FMI members was to be expected, partly because of scale mismatch as the two reanalysis datasets represent the meteorological average of a larger area than that covered with the in situ meteorological measurements. More critically, the answer to question 2 is also "yes" and the artifacts mentioned in the question were found to assume multiple forms:

1) At times, performance metrics of the NCEP and WFDEI members suggested that they performed well, but a closer inspection of the time series and a focus on seasonal performance revealed that they did not. For example, Fig. 4b suggested that statistical errors in

NCEP T_s were low. While this, in theory, is correct, Fig. 4a showed that the top range of modeled winter soil temperatures (obtained by the NCEP member) was above 0°C and Fig. 8 showed that seasonal differences in NCEP T_s were largely underestimated. Failing to identify such seemingly small differences in soil temperatures at high latitude could have important implications in climate studies, which rely heavily on aggregated metrics, because the response of the soil thermal regime to slight changes in temperatures around the melting point far exceeds the scale of the temperature change itself. Equally, the decreased performance in sensible heat fluxes of the WFDEI members in Fig. 9 compared to Fig. 8 warrants further investigations because of the expected shortening of the shoulder seasons under a warming climate (Räisänen 2008; Lawrence and Slater 2010). Such errors are unrelated to scale; the WFDEI and NCEP members differed considerably and both covered the same area.

2) The ability of the model to reproduce the snow depth and water equivalent had a considerable effect on all of the other evaluated model outputs. In other words, most of the errors discussed in the previous sections were related to snow and, more precisely, to precipitation amount. While this should not come as a surprise given that the site is situated in the boreal ecozone, the benchmarking of JULES (Blyth et al. 2011) attributed some of the errors in peak flow in snow-covered sites to model discrepancies; Hancock et al. (2014) suggested that, had the benchmarking evaluated snow amount and/or cover at the sites, some of these errors would instead have been attributed to errors in precipitation in the reanalysis data.
3) In the course of this research, the authors found that three switches, which were added to the input files between JULES 3.0 and 3.3 but were not highlighted in the JULES user guide, considerably affected results: one caused a reduction in H and increase in LE during snowmelt at the forest site and two led to premature snowmelt (details in the appendix). All published studies using these versions of JULES in snow-covered regions are likely to have used suboptimal model setting; had this experimental study not been so highly controlled and one of the authors not been involved in the development of snow processes representation in earlier versions of JULES, the reported performance of the ensemble would have differed considerably.

Over the past few years, the land surface community has concentrated much effort to improve and standardize the way LSMs are evaluated. Benchmarking is becoming a priority and proposed frameworks generally advocate that hydrological and biogeochemical cycles be evaluated across broad temporal and spatial scales (e.g., Williams et al. 2009; Abramowitz 2012; Luo et al. 2012). While these frameworks are laudable, their implementation does not consider pragmatic considerations that, if ignored, could diminish the impact of benchmarking. Such considerations include the choice of the modelers themselves (e.g., their research background and their familiarity with the model and its process representations and hundreds of switches and parameters), allocation of research funding to different teams at different times, time constraints, etc.

Evaluation or benchmarking of a global-scale model is problematic because issues of scale remain unsolved. Zhao et al. (2012), who evaluated meteorological variables from four reanalysis datasets at different spatial scales, found that differences in spatial resolution between datasets were significantly smaller than those caused by differences between the datasets, thus suggesting that scale mismatch was not the most significant issue. Nevertheless, there is currently no method to evaluate large-scale distributed model simulations against distributed measurements as rigorous as point-to-point evaluation. LSMs are either forced by meteorological measurements (e.g., Blyth et al. 2011; Best et al. 2015) and evaluated against multiple-site data representative of specific ecozones or are forced by reanalysis data and evaluated against satellite products. These products are themselves often evaluated against points measurements (e.g., Takala et al. 2011; Loew and Schlenz 2011; Yang et al. 2015) or reanalysis data (e.g., Dorigo et al. 2010) or are assimilated with reanalysis data (e.g., Zhao et al. 2006); reanalysis datasets too are evaluated against local data (Weedon et al. 2014). As a consequence, the methods used in this study are still the best currently available to upscale the evaluation process and to avoid circular (i.e., evaluate model performance when using reanalysis data against satellite products that assimilate said data) evaluation.

This study quantified the effect of meteorological and ancillary data and temporal averaging on model performance and scrutinized the modeler's choice of performance metrics on interpretation of the results. The surface processes investigated, although evaluated here at a single site, are at the core of land surface–climate feedbacks (e.g., snow albedo, soil–carbon). Yet, in benchmarking exercises or multiple-site evaluation of LSMs, there are often either no high-latitude sites evaluated (e.g., Abramowitz et al. 2008; Best et al. 2015) or the influence of the accurate modeled representation of snow on the evaluated processes in snow-covered sites is omitted (e.g., Williams et al. 2009; Blyth et al. 2010, 2011; Slevin et al. 2015). In view of the importance of snow on the range of the results obtained with the same model, let alone

identical simulations using different temporal averaging, it is recommended that systematic evaluation, quantification of errors, and uncertainties in snow-covered regions be incorporated in benchmarking frameworks.

Acknowledgments. This research was funded by the EU LIFE+ MONIMET (LIFE12 ENV/FI/000409) and ESA SMOS+ Innovation Permafrost (ESRIN 4000 105184/12/I-BG) projects. The authors thank Ari Aaltonen, Timo Ryyppö, and Riika Ylitalo (FMI, Finland) for providing meteorological and evaluation data from the FMI-ARC in Sodankylä; Matt Pryor (Met Office) and Douglas Clark (Centre for Ecology and Hydrology, United Kingdom) for their support in debugging JULES; Richard Essery (University of Edinburgh, United Kingdom) and Graham Weedon (Met Office) for constructive comments on an earlier version of the manuscript. JULES is freely available to researchers for non-commercial use and can be requested from https://jules.jchmr.org/software-and-documentation. The FMI meteorological and evaluation data are available on http://litdb.fmi.fi/.

APPENDIX

Description of Problematic End-User Switches

First, a new end-user switch (l_snowdep_surf) was added in JULES 3.0 to allow turbulent fluxes to be calculated as if all the snow were on the canopy, even when the snow was on the ground. In all previous versions of JULES, if partitioning of snow was selected (by setting can_mod = 4), the calculation of turbulent fluxes implicitly accounted for whether snow was on the canopy or not. In all JULES 3.*n*, where *n* is a subversion number, the default settings allow partitioning of snow between canopy and ground but assume all snow on the canopy for calculation of the turbulent fluxes. There is no mention in the manual of the reduced functionality of can_mod = 4.

Second, a choice between spectral and all-band albedo schemes is available to JULES users. Until JULES 3.3, if the spectral albedo scheme was selected, JULES calculated a spectral snow albedo. In 3.3, a new switch was added to allow users to provide observed snow albedo (l_albedo_obs). As a consequence, another switch was added to choose whether, if spectral albedo is selected, spectral snow albedo is also required (l_snow_albedo). When JULES 3.3, 3.4, and 3.4.1 were released, the default settings switched off "l_snow_albedo" and, although the manual mentioned the new switch in its release notes, a description of l_snow_albedo did not feature as a new switch in the description of the input files. These omissions were corrected retrospectively in the online documentation for JULES 3.3. to 3.4.1. following identification of the issue in the course of this study.

REFERENCES

Abramowitz, G., 2012: Towards a public, standardized, diagnostic benchmarking system for land surface models. *Geosci. Model Dev.*, **5**, 819–827, doi:10.5194/gmd-5-819-2012.

——, R. Leuning, M. Clark, and A. Pitman, 2008: Evaluating the performance of land surface models. *J. Climate*, **21**, 5468–5481, doi:10.1175/2008JCLI2378.1.

Best, M. J., and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES), model description—Part 1: Energy and water flux. *Geosci. Model Dev.*, **4**, 677–699, doi:10.5194/gmd-4-677-2011.

——, and Coauthors, 2015: The plumbing of land surface models: Benchmarking model performance. *J. Hydrometeor.*, **16**, 1425–1442, doi:10.1175/JHM-D-14-0158.1.

Blyth, E., J. Gash, A. Lloyd, M. Pryor, G. Weedon, and J. Shuttleworth, 2010: Evaluating the JULES land surface model energy fluxes using FLUXNET data. *J. Hydrometeor.*, **11**, 509–519, doi:10.1175/2009JHM1183.1.

——, D. B. Clark, R. Ellis, C. Huntingford, S. Los, M. Pryor, M. Best, and S. Sitch, 2011: A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale. *Geosci. Model Dev.*, **4**, 255–269, doi:10.5194/gmd-4-255-2011.

Bowling, L., and Coauthors, 2003: Simulation of high-latitude hydrological processes in the Torne–Kalix basin: PILPS phase 2(e): 1: Experiment description and summary intercomparisons. *Global Planet. Change*, **38**, 1–30, doi:10.1016/S0921-8181(03)00003-1.

Clapp, R. B., and G. M. Hornberger, 1978: Empirical equations for some soil hydraulic properties. *Water Resour. Res.*, **14**, 601–604, doi:10.1029/WR014i004p00601.

Clark, D. B., and Coauthors, 2011: The Joint UK Land Environment Simulator (JULES), model description—Part 2: Carbon fluxes and vegetation. *Geosci. Model Dev.*, **4**, 701–722, doi:10.5194/gmd-4-701-2011.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay, 2008: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resour. Res.*, **44**, W00B02, doi:10.1029/2007WR006735.

——, D. Kavetski, and F. Fenicia, 2011: Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.*, **47**, W09301, doi:10.1029/2010WR009827.

Cosby, B. J., G. M. Hornberger, R. B. Clapp, and T. R. Ginn, 1984: A statistical exploration of the relationships of soil-moisture characteristics to the physical properties of soils. *Water Resour. Res.*, **20**, 682–690, doi:10.1029/WR020i006p00682.

Deardorff, J., 1978: Efficient prediction of ground surface temperature and moisture, with inclusion of a layer of vegetation. *J. Geophys. Res.*, **83**, 1889–1903, doi:10.1029/JC083iC04p01889.

Decker, M., M. A. Brunke, Z. Wang, K. Sakaguchi, X. Zeng, and M. G. Bosilovich, 2012: Evaluation of the reanalysis products from GSFC, NCEP, and ECMWF using flux tower observations. *J. Climate*, **25**, 1916–1944, doi:10.1175/JCLI-D-11-00004.1.

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.

Déqué, M., and Coauthors, 2007: An intercomparison of regional climate simulations for Europe: Assessing uncertainties in model projections. *Climatic Change*, **81**, 53–70, doi:10.1007/s10584-006-9228-x.

Dorigo, W. A., K. Scipal, R. M. Parinussa, Y. Y. Liu, W. Wagner, R. A. M. de Jeu, and V. Naeimi, 2010: Error characterisation of global active and passive microwave soil moisture datasets. *Hydrol. Earth Syst. Sci.*, **14**, 2605–2616, doi:10.5194/hess-14-2605-2010.

Essery, R., 2013: Large-scale simulations of snow albedo masking by forests. *Geophys. Res. Lett.*, **40**, 5521–5525, doi:10.1002/grl.51008.

——, J. Pomeroy, J. Parviainen, and P. Storck, 2003: Sublimation of snow from coniferous forests in a climate model. *J. Climate*, **16**, 1855–1864, doi:10.1175/1520-0442(2003)016<1855:SOSFCF>2.0.CO;2.

——, and Coauthors, 2009: An evaluation of forest snow process simulations. *Bull. Amer. Meteor. Soc.*, **90**, 1120–1135, doi:10.1175/2009BAMS2629.1.

——, S. Morin, Y. Lejeune, and C. B. Ménard, 2013: A comparison of 1701 snow models using observations from an alpine site. *Adv. Water Resour.*, **55**, 131–148, doi:10.1016/j.advwatres.2012.07.013.

Etchevers, P., and Coauthors, 2004: Validation of the energy budget of an alpine snowpack simulated by several snow models (SnowMIP project). *Ann. Glaciol.*, **38**, 150–158, doi:10.3189/172756404781814825.

European Environment Agency, 2006: CORINE land cover, CLC2006-maankytt/maanpeite (25m) -paikkatietoaineisto (rasteri). Accessed 20 May 2014. [Available online at http://wwwd3.ymparisto.fi/d3/Static_rs/spesific/corinelandcover.html.]

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972.

Hancock, S., B. Huntley, R. Ellis, and B. Baxter, 2014: Biases in reanalysis snowfall found by comparing the JULES land surface model to GlobSnow. *J. Climate*, **27**, 624–632, doi:10.1175/JCLI-D-13-00382.1.

Houldcroft, C. J., W. M. Grey, M. Barnsley, C. M. Taylor, S. O. Los, and P. R. J. North, 2009: New vegetation albedo parameters and global fields of soil background albedo derived from MODIS for use in a climate model. *J. Hydrometeor.*, **10**, 183–198, doi:10.1175/2008JHM1021.1.

JCHMR, 2013: Joint UK Land Environment Simulator User Guide. Accessed 2 May 2014. [Available online at http://www.jchmr.org/jules/documentation/user_guide/vn3.4.1/.]

Jolliff, J. K., J. C. Kindle, I. Shulman, B. Penta, M. A. M. Friedrichs, R. Helber, and R. A. Arnone, 2009: Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *J. Mar. Syst.*, **76**, 64–82, doi:10.1016/j.jmarsys.2008.05.014.

Lawrence, D., and A. Slater, 2010: The contribution of snow condition trends to future ground climate. *Climate Dyn.*, **34**, 969–981, doi:10.1007/s00382-009-0537-4.

Loew, A., and F. Schlenz, 2011: A dynamic approach for evaluating coarse scale satellite soil moisture products. *Hydrol. Earth Syst. Sci.*, **15**, 75–90, doi:10.5194/hess-15-75-2011.

Luo, Y. Q., and Coauthors, 2012: A framework for benchmarking land models. *Biogeosciences*, **9**, 3857–3874, doi:10.5194/bg-9-3857-2012.

Manninen, T., L. Korhonen, P. Voipio, P. Lahtinen, and P. Stenberg, 2012: Airborne estimation of boreal forest LAI in winter conditions: A test using summer and winter ground truth. *IEEE Trans. Geosci. Remote Sens.*, **50**, 68–74, doi:10.1109/TGRS.2011.2173939.

McGuire, A., and Coauthors, 2001: Carbon balance of the terrestrial biosphere in the twentieth century: Analyses of CO_2, climate and land use effects with four process-based ecosystem models. *Global Biogeochem. Cycles*, **15**, 183–206, doi:10.1029/2000GB001298.

Moriasi, D. N., J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, 2007: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE*, **50**, 885–900, doi:10.13031/2013.23153.

Pitman, A. J., 2003: The evolution of, and revolution in, land surface schemes designed for climate models. *Int. J. Climatol.*, **23**, 479–510, doi:10.1002/joc.893.

Pomeroy, J. W., D. M. Gray, T. Brown, N. R. Hedstrom, W. L. Quinton, R. J. Granger, and S. K. Carey, 2007: The cold regions hydrological process representation and model: A platform for basing model structure on physical evidence. *Hydrol. Processes*, **21**, 2650–2667, doi:10.1002/hyp.6787.

Räisänen, J., 2008: Warmer climate: Less or more snow? *Climate Dyn.*, **30**, 307–319, doi:10.1007/s00382-007-0289-y.

Reid, T., R. Essery, N. Rutter, and M. King, 2013: Data-driven modelling of shortwave radiation transfer to snow through boreal birch and conifer canopies. *Hydrol. Processes*, **28**, 2987–3007, doi:10.1002/hyp.9849.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks, 2010: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resour. Res.*, **46**, W05521, doi:10.1029/2009WR008328.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, doi:10.1175/2010BAMS3001.1.

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Sheng, J., and F. Zwiers, 1998: An improved scheme for time-dependent boundary conditions in atmospheric general circulation models. *Climate Dyn.*, **14**, 609–613, doi:10.1007/s003820050244.

Slater, A., and Coauthors, 2001: The representation of snow in land surface schemes: Results from PILPS2(d). *J. Hydrometeor.*, **2**, 7–25, doi:10.1175/1525-7541(2001)002<0007:TROSIL>2.0.CO;2.

Slevin, D., S. Tett, and M. Williams, 2015: Multi-site evaluation of the JULES land surface model using global and local data. *Geosci. Model Dev.*, **8**, 295–316, doi:10.5194/gmd-8-295-2015.

Takala, M., K. Luojus, J. Pulliainen, C. Derksen, J. Lemmetyinen, J.-P. Kärnä, J. Koskinen, and B. Bojkov, 2011: Estimating Northern Hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements. *Remote Sens. Environ.*, **115**, 3517–3529, doi:10.1016/j.rse.2011.08.014.

Taylor, K., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, doi:10.1029/2000JD900719.

Vincent, L. A., X. Zhang, B. R. Bonsal, and W. D. Hogg, 2002: Homogenization of daily temperatures over Canada. *J. Climate*, **15**, 1322–1334, doi:10.1175/1520-0442(2002)015<1322:HODTOC>2.0.CO;2.

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.

Williams, M., and Coauthors, 2009: Improving land surface models with FLUXNET data. *Biogeosciences*, **6**, 1341–1359, doi:10.5194/bg-6-1341-2009.

Wood, E., and Coauthors, 1998: The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) Phase 2(c) Red–Arkansas River basin experiment: 1. Experiment description and summary intercomparisons. *Global Planet. Change*, **19**, 115–135, doi:10.1016/S0921-8181(98)00044-7.

Wunderlich, T., 1972: Heat and mass transfer between a water surface and the atmosphere. Water Resources Research Laboratory Rep. 0-6803 14, Tennessee Valley Authority, 270 pp.

Yang, J., L. Jiang, C. B. Ménard, K. Luojus, J. Lemmetyinen, and J. Pulliainen, 2015: Evaluation of snow products over the Tibetan Plateau. *Hydrol. Processes*, **29**, 3247–3260, doi:10.1002/hyp.10427.

Zhao, M., S. W. Running, and R. R. Nemani, 2006: Sensitivity of Moderate Resolution Imaging Spectroradiometer (MODIS) terrestrial primary production to the accuracy of meteorological reanalyses. *J. Geophys. Res.*, **111**, G01002, doi:10.1029/2004JG000004.

Zhao, Y., and Coauthors, 2012: How errors on meteorological variables impact simulated ecosystem fluxes: A case study for six French sites. *Biogeosciences*, **9**, 2537–2564, doi:10.5194/bg-9-2537-2012.