

Verification of land-atmosphere coupling in forecast models, reanalyses and land surface models using flux site observations

Paul A. Dirmeyer^{1*}, Liang Chen¹, Jiexia Wu¹, Chul-Su Shin¹, Bohua Huang¹, Benjamin A. Cash¹, Michael G. Bosilovich², Sarith Mahanama², Randal D. Koster², Joseph A. Santanello², Michael B. Ek³, Gianpaolo Balsamo⁴, Emanuel Dutra⁵, and D. M. Lawrence⁶

¹Center for Ocean-Land-Atmosphere Studies, George Mason University

²NASA / Goddard Space Flight Center

³NOAA / National Centers for Environmental Prediction / Environmental Modeling Center

⁴European Centre for Medium-range Weather Forecasts

⁵Instituto Dom Luiz, Faculdade de Ciências, Universidade de Lisboa⁶National Center for Atmospheric Research

*Corresponding Author:

Paul A. Dirmeyer
Center for Ocean-Land-Atmosphere Studies
George Mason University
4400 University Drive, Mail Stop: 6C5
Fairfax, Virginia 22030 USA
pdirmeye@gmu.edu

Submitted to: *Journal of Hydrometeorology*

Abstract:

We confront four model systems in three configurations (LSM, LSM+GCM, and reanalysis) with global flux tower observations to validate states, surface fluxes, and coupling indices between land and atmosphere. Models clearly under-represent the feedback of surface fluxes on boundary layer properties (the atmospheric leg of land-atmosphere coupling), and may over-represent the connection between soil moisture and surface fluxes (the terrestrial leg). Models generally under-represent spatial and temporal variability relative to observations, which is at least partially an artifact of the differences in spatial scale between model grid boxes and flux tower footprints. All models bias high in near-surface humidity and downward shortwave radiation, struggle to represent precipitation accurately, and show serious problems in reproducing surface albedos. These errors create challenges for models to partition surface energy properly and errors are traceable through the surface energy and water cycles. The spatial distribution of the amplitude and phase of annual cycles (first harmonic) are generally well reproduced, but the biases in means tend to reflect in these amplitudes. Interannual variability is also a challenge for models to reproduce. Our analysis illuminates targets for coupled land-atmosphere model development, as well as the value of long-term globally-distributed observational monitoring.

1. Introduction

Many LSMs were developed and pressed into service during the 1980s and 1990s to provide lower boundary conditions for the atmospheric GCMs used in climate and weather simulation and prediction (Santanello et al. 2017). This occurred at a time when observations of key land surface variables, and the coupled processes that link the water and energy cycles between the land and atmosphere, were extremely limited. As a result, performance of coupled LSM-GCM systems has been sub-optimal (Dirmeyer et al. 2017).

The necessary observational data sets for validation are only recently becoming available; datasets that combine co-located measurements of land surface states, surface fluxes, near-surface meteorology, and properties of the atmospheric column. Early field campaigns (e.g., Sellers et al. 1992, 1995; Famiglietti et al. 1999; Jackson and Hsu 2001; Andreae 2002) provided observations that helped advance theory and model parameterization development, but their short periods of operation meant collected data provided limited sampling of the phase-space of land-atmosphere interactions, rarely quantifying interannual variability. In the mid-1990s, networks of observing stations began to be established and maintained, providing long-term data sets. A growing number of soil moisture monitoring networks have been established. Their data have been collated, homogenized and standardized by two separate efforts (Dorigo et al. 2011, 2013, 2017; Quiring et al. 2016). Those data sets were used by Dirmeyer et al. (2016) in a first-of-its-kind multi-model multi-configuration assessment of soil moisture simulation fidelity.

Simultaneously, efforts began in the ecological community to collect surface flux data over a variety of biomes (FLUXNET; Baldocchi et al 2001). Over time, in consultation with interested scientific communities, FLUXNET expanded their instrumentation suite to measure soil moisture, ground heat flux, and four-component radiation, allowing detailed

closure of the surface energy balance. Rigid standards for data formatting and dissemination within and across regional networks was lacking, so a global standardized and quality-controlled subset of data from many FLUXNET sites ³ (<http://www.fluxdata.org>) covering multiple links in the coupled land-atmosphere process chain (Santanello et al. 2011). The La Thuile data set enabled a greater degree of model validation (e.g., Williams et al. 2009; Bonan et al. 2012; Boussetta et al. 2013; Melaas et al. 2013; Balzarolo et al. 2014; Purdy et al. 2016).

In this study, we employ the updated FLUXNET2015 synthesis data set, (Pastorello et al. 2017) expanding the multi-model multi-configuration study of soil moisture simulations in Dirmeyer et al. (2016) to a global assessment of surface energy and water balance simulations, and basic metrics of land-atmosphere coupling. Section 2 describes the observational data and models examined. The next three sections present validations of model annual means, annual cycles, and coupling metrics. We then discuss some of the pathological model behaviors that emerge from the analysis and present conclusions. Throughout the paper we present synthesis figures. Detailed scatter plots showing results across all FLUXNET2015 sites for each model are consigned to the Supplement.

2. Data and Models

The range of dates of data varies considerably among model simulations, and also between individual observational sites. We analyze spatial variability and compare only climatologies (annual means or mean annual cycles) in order to minimize the effect of such asynchronicities, and present a quantification of interannual variability. It is not the intent of this study to validate model simulations of specific events, but rather their overall coupled land-atmosphere behavior. Note also that many coupling metrics, including those used here,

can be calculated for LSMs from a combination of forcing and model output, even though the LSMs are not coupled to GCMs.

2.1 Observed data

In situ measurements of near surface meteorological variables, surface fluxes and soil moisture used for model validation come from the November 2016 version of the FLUXNET2015 station data set. Daily, monthly and yearly data have been used; processing of the meteorological, radiation, heat flux and surface hydrologic data including gap-filling are described by Reichstein et al. (2005) and Vuichard and Papale (2015). Only the Tier 1 (open access) data are used in this study (see Table S1 for a complete list of sites) Figure 1 shows the spatial distribution of sites and some of the key characteristics regarding data availability. 166 sites provide 1242 site-years of data, but coverage is concentrated in the mid-latitudes and particular underrepresentation in the tropics.

The variables processed for this analysis include surface pressure, near surface air temperature and vapor pressure deficit, precipitation, four-component and net radiation, surface sensible and latent heat fluxes (gap-filled following the method of Reichstein et al. 2005 and energy balance closure-corrected) and soil water content measured at the first (shallowest) sensor. There is no consolidated information on the depth of the shallowest sensor across all sites, but typically it is at 5cm or 10cm below the surface. Vapor pressure deficit is converted to specific humidity using the Clausius-Clapeyron relationship. We have used the provided FLUXNET2015 data at the corresponding time intervals for each calculation: yearly data for annual means, monthly data for annual cycles, and daily data for calculating coupling indices.

In addition, we examine a number of gridded global precipitation products for comparison to FLUXNET2015 sites. These are listed in Table S2.

2.2 Model systems

Four global modeling systems are evaluated; two from operational forecast centers and two that are primarily used for research. The operational systems are from the U.S. National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-range Weather Forecasts (ECMWF). The research systems are from the U.S. National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO) and the U.S. National Center for Atmospheric Research (NCAR).

Table 1 summarizes the model components and configurations. Generally, each modeling system is interrogated in three different configurations: 1) LSM only (offline), driven by gridded observationally-based meteorological analyses including downward radiation; 2) LSM coupled to GCM in a *free-running* mode where the coupled system evolves unconstrained after initialization; 3) Reanalysis, where the coupled LSM and GCM are constrained by data assimilation at diurnal or sub-diurnal increments to represent the actual historical evolution of state variables. The NCAR model system does not have an associated reanalysis, so to keep the four-by-three matrix filled, two different reanalyses from GMAO are included. Note that when the coordinates for a FLUXNET2015 site lie within a © « Ÿ ĭ " - ocean grid cell, it is excluded from comparisons for that model. Thus, the number of stations compared vary from model to model depending on resolution and the land-sea mask.

2.2.1 NCEP

Data for the offline configuration comes from an author-produced simulation using Noah LSM version 2.7.1 (Ek et al., 2003, Mitchell, 2005) driven by 3-hourly gridded meteorological data from the Terrestrial Hydrology Research Group at Princeton University (Sheffield et al., 2006). The free-running coupled land-atmosphere simulation consists of a subset of 48 years

from a 420 year long current climate simulation of CFSv2 initialized in 1980 (Shukla et al. 2017). The coupled simulation is unique among the model systems in that it also includes a coupled ocean component. However, this should have very little effect on the local coupled land-atmosphere behavior of the model. Years 2101-2148 of the simulation are used, but the calendar dates have no real meaning in a fully coupled climate model so far from the initial state, wherein attributes such as atmospheric composition, solar intensity, orbital parameters, etc., are held constant at late 20th century values. The latest NCEP reanalysis is also examined (CFSR; Saha et al. 2010), which combines a global land data assimilation system derived from the NASA Land Information System (LIS; Peters-Lidard et al., 2007), driven by a blended global precipitation analysis (Xie and Arkin 1997; Xie et al. 2007), used to update the coupled analysis cycle once per day over the period 1979-2009.

2.2.2 GMAO

Two reanalyses are included for GMAO; version 1 and version 2 of the Modern-Era Retrospective Analysis for Research and Applications (MERRA; Rienecker et al. 2011, Reichle et al. 2017a). MERRA data cover the period 1980-2015. MERRA-2 is the current state-of-the-art reanalysis covering 1980-2015 (Molod et al. 2015, Gelaro et al. 2017), and is the source of most of the meteorological forcing data for the offline simulation of the Catchment LSM v25 C05 (GMAO 2015a,b). As part of the MERRA-2 reanalysis, the GCM-generated precipitation is corrected with observations-based precipitation before it reaches the land surface (Reichle et al. 2017b); the reanalysis meteorological fields thus feel the observed precipitation rates indirectly through the surface fluxes. Additionally, a global 36-year offline Catchment simulation on the MERRA grid and a 16-year coupled GEOS5-Catchment simulation at half-degree resolution with prescribed observed SSTs were generated for this comparison.

2.2.3 NCAR

There is no operational reanalysis produced with the NCAR Community Earth System Model (CESM). However, CESM is widely used for research in the academic community, and we have generated offline and coupled simulations for this comparison. The offline simulation uses version 4.5 of the Community Land Model (CLM; Lawrence et al. 2011) driven with forcing spanning 1991-2010 from version 4 of the blended and gap-filled CRUNCEP (Viovy 2013) 0.5° data set (available at: <https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm4.CRUNCEP.v4.html>) aggregated to the nominal 1° GCM resolution. A simulation with CLM4.5 coupled to CAM4 in CESM1.2.2 has been produced spanning 1991-2014 with specified climatological SSTs.

2.2.4 ECMWF

The offline simulation from ECMWF is with Cycle 43R1 of the Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land (HTESSEL) run at ~16km resolution based on a cubic octahedral global grid (TCO639) for the period 1979-2015. This offline simulation follows ERA-Interim/land configurations closely (see Balsamo et al. 2015), forced by ERA-Interim meteorology and fluxes with **an altitude correction applied to temperature, humidity and surface pressure**. This offline simulation is used to initialize the land state of the operational ECMWF hindcasts. The coupled simulation comes from the Athena Project (Kinter et al. 2013) for 1961-2007 where an older version of HTESSEL is coupled to IFS Cycle 32R3 at a similarly high native horizontal resolution and specified observed SSTs, but the data has been post-processed to a 1.125° uniform grid. ERA-Interim (Dee et al. 2011), spanning 1979-2015, provides the reanalysis configuration of data for the comparison, which used TESSEL prior to hydrology upgrades.

3. Annual Means

The comparison of models to FLUXNET2015 observations of annual means amounts to an assessment of model ability to reproduce global spatial patterns (within the limitations of the uneven distribution of station locations) of the variables. For the offline LSM simulations, meteorological forcing data are specified from gridded data sets, so their correlation to FLUXNET2015 observations is not a pure reflection of model performance as the forcing data constrain LSM behavior. Similarly, for the reanalysis products, performance reflects a combination of model characteristics, data assimilation techniques and the distribution and quality of the data assimilated. Assimilation of observational data constrains the coupled land-atmosphere model behavior to some degree. While the free-running model simulations provide an unabridged assessment of model performance, results from the other modes of simulation are nevertheless enlightening.

As an indicator of observational uncertainty and the impact of comparing model grid box values to field sites, we first note how a number of gridded observational precipitation products and the reanalyses validate against precipitation measurements at FLUXNET2015 locations. Figure 2 shows mean (dots) and span (whiskers) of annual precipitation totals, where the abscissa always corresponds to measurements from the FLUXNET2015 sites. For most sites, the observational products (top two rows of Fig. 2) cover the entire time span of FLUXNET2015 observations (see Table S2 for details). All reanalyses (bottom row of Fig. 2) except CFSR span the FLUXNET2015 period. Several statistics of spatial agreement are produced: moment correlation coefficient (r_p), Spearman's rank correlation coefficient (r_s), root mean square error (RMSE), slope of the best-fit linear regression of Y on X (Slope) and the fraction of total stations $N_{\text{span}}/N_{\text{total}}$ where the span of the individual annual totals from the gridded products (vertical whiskers) overlap the span

from FLUXNET2015 sites (horizontal whiskers). The last statistic tests the possibility that the FLUXNET2015 observations and gridded estimates do not come from distinct populations, i.e. their ranges overlap.

Estimates from gridded observational data sets, which range in spatial resolution from 0.25° (MSWEP, TRMM) to 2.5° (GPCP), provide a plausible upper bound to the accuracy we could expect from gridded Earth system models. For the 166 (or fewer) FLUXNET2015 sites compared, which admittedly represent a rather uneven sampling of global terrestrial precipitation, three observational products score at the top: MSWEP, CPC-Uni and U.Del. The correlation of nearly 0.8, a rank correlation between 0.8-0.9, and the highest number of stations whose ranges span the diagonal $X=Y$ line. The lower limit for RMSE across these sites is about 240mm. Note that all gridded products underestimate the slope, indicating the inability of large area averages to resolve local variations in average precipitation.

MERRA-2 performs on par with the best gridded observed products, namely because it reports a bias corrected precipitation that is used as part of the assimilation process instead of model-generated precipitation as an input to the LSM (Reichle and Liu 2014). Thus, it is effectively another gridded observational data set for precipitation. Figure S1 compares the precipitation predicted by the model physical parameterizations in MERRA-2 alongside the corrected version in the same fashion as Fig 2. The correction greatly reduces bias, cuts RMSE by one third, slightly improves spatial correlations, and increases the number of stations spanning the diagonal by 28%. CFSR significantly underperforms other reanalyses at FLUXNET2015 locations.

Precipitation is among the most difficult quantities for models to simulate. We expect among near surface meteorological variables the lowest correlations and largest coefficient

of variation for precipitation. It also has many observationally-based data sets to choose from, providing a robust estimate of skill to be expected from comparing point measurements to gridded data sets. Figure 2 provides generous thresholds, particularly for correlations, to keep in mind when assessing model simulations of the terms of the surface water and energy balance. As shown below, correlations of 0.7-0.8 are a challenge for models to attain for precipitation, as well as some other water and energy budget terms.

Among near surface meteorology (e.g., temperature and specific humidity) and downward surface fluxes (including shortwave and longwave radiation), precipitation has the greatest small-scale variability on monthly to annual time scales, and is thus the most difficult land surface variable to simulate. Figures S2-S6 show the scatters and statistics for the models listed in Table 1 for these five variables. Here, the restriction that the years of the models match those at each FLUXNET2015 site is lifted, and the climatologies of the complete data sets are compared. Not surprisingly, the global distribution of annual mean temperature is very well reproduced by the models (Fig. S2), with 88-96% of the observed variance explained. Observed specific humidity is only slightly less well correlated among the models (Fig. S3), but there is a consistent positive bias relative to FLUXNET2015 measurements. Patterns of annual mean downward radiation (Figs. S4 and S5) are well simulated, with a tendency for a slight negative bias in longwave radiation (Fig. S5), and a stronger positive bias in shortwave radiation across models (Fig. S4), consistent with other assessments of model shortwave errors that depend on GCM radiative transfer parameterizations (cf. Slater 2016). Precipitation shows the least agreement; note the bottom row of Fig. S6 is not identical to that of Fig. 2 because the years compared differ. Nevertheless, the results are similar. We can consider MERRA-2 as representing the upper limit of comparison for annual precipitation when the periods do not match between models and observations. Offline Catchment actually performs slightly better

than MERRA-2, and CFSv2 is generally the poorest performing model system in the set. Free-running climate models understandably perform worse than either reanalyses or offline LSM simulations, as they are least constrained by observational data. In the case of CFSv2, there are essentially no constraints within the Earth system as an ocean model is coupled; other free-running simulations have specified SSTs.

Precipitation is a major source of error at the land surface, but so are elements of the radiation budget. We employ Taylor diagrams to synthesize the statistics of correlation across FLUXNET2015 sites; RMSE and standard deviation are normalized by observed values. Figure 3 shows the global distribution of annual mean downward radiation terms is well simulated across all model configurations, with downward shortwave radiation performing slightly better than downward longwave radiation. Recall for the LSM-only models, downward radiation is an input forcing, and the quality of those data sets can vary significantly (Slater 2016). However, the distribution of upward shortwave radiation is rather poorly simulated, with the NCEP models showing the worst correlations, and the NCAR models the best (yet explaining less than half of the variance). There is also a strong tendency to under-represent the spatial variability (normalized standard deviations less than 1) of downward shortwave radiation. This degrades simulation of net radiation, which has consistently lower correlations than downward radiation terms, yet uniformly better than upward shortwave radiation. The overlap of the spans of annual mean values from models and observations (size of the dots) generally decrease from shortwave down to longwave down to shortwave up.

Figure 3 implies discrepancies in the representation of surface albedo across models at FLUXNET2015 sites. We show a Taylor diagram for calculated albedo in Fig. 4. As there are many sites at relatively high northern latitudes that experience snow cover for some part of

the year, snow albedo could specifically be a problem. However, a plot of only the JJA albedo verification shows boreal summer generally has even lower fidelity, and systematically low spatial variability, compared to the annual mean. The overlap between the spans of annual mean albedos range among the models from 16% to 38% of FLUXNET2015 sites, but for JJA they span only 13-24%.

The low variability could be explained by the fact that most LSMs, whether stand-alone or coupled, have a simple parameterization of albedo based on properties of a small number of vegetation and soil types, often specified as a climatological seasonal cycle. CLM actually calculates surface albedo based on a number of properties including vegetation density and zenith angle of the sun, which may lead to the somewhat better performance of the NCAR models. As described later, the offline NCEP LSM (identified as NL) specifies a multi-year satellite-derived monthly green vegetation fraction as a boundary condition that appears in Fig. 4 to enhance variability, while its positive biases have been noted by Xia et al. (2012). Furthermore, discrepancies between grid box average albedo and local conditions at field sites, including the effect of vegetation differences and soil moisture on albedo (Zaitchik et al. 2013), are representative of the larger problem of representing the spatial and temporal variability of albedo. Nevertheless, such discrepancies lead to a degradation in the representation of surface available energy that is partitioned between sensible, latent and ground heat fluxes. Even an otherwise perfect LSM could not produce the right values of these fluxes if net radiation is incorrect. Coupled with errors in precipitation, which affect available soil moisture and thus Bowen ratios, LSMs are at a compounded disadvantage in simulating the surface water and energy budget terms.

In Fig. 5 we correlate across the stations the mean errors in key water and energy cycle quantities and present a schematic representation of the relative coupling or connectedness

exhibited between terms. This also suggests how errors in the simulation or specification of one term can propagate to others through the land-atmosphere coupling process chain (cf. Santanello et al. 2011). r_s is generally larger than r_p because it does not overemphasize outliers, thus is used for this comparison. Ratios show the fraction of models with correlations at the 90% confidence level, and p-values are based on the average correlation across models. Note the number of included stations varies depending on the availability of observed data (recall from Fig. 1 that a number of FLUXNET2015 sites do not allow for albedo estimations) and among models depending on whether the corresponding grid box is water or land. Furthermore, the data saved from the free-running ECMWF model simulations (EC) do not allow for estimation of albedo, so 11 models are compared for albedo.

Unsurprisingly, we find surface net radiation errors correlate strongly to albedo errors, with 11 of 11 models registering significant correlations (two-tailed p-values < 0.05) and the multi-model average correlation across 114-118 sites has a p-value of 4×10^{-7} . For net radiation versus precipitation, only 2 of 12 models (CL and M1) show significant correlation across 144-151 sites and $p=0.55$ for the multi-model average, so no direct arrow is drawn in Fig. 5. Note that precipitation errors arise not only from misrepresentation of land-atmosphere interactions, but also from the parameterization of dynamic and thermodynamic processes (so-œš " " j Ÿ " © « Ÿ j " " ¬ ¤ µ " ¥ œ " ¥ ª " ° ¤ j " fi

FLUXNET2015 reports both raw and Bowen-ratio corrected heat fluxes. Corrected fluxes are available at fewer than 100 of the sites (two-tailed $p=0.05$ for correlations 0.2, compared to 0.16 for the full set of sites), but generally correspond better to the models than uncorrected fluxes, which do not close the surface energy balance (cf. Figs. S9-S12). Regardless, the same story emerges with either set of fluxes: precipitation errors correlate

873

874 Figure 6: As in Fig. 3 for the magnitude of the annual cycle (first harmonic calculated from
875 monthly means) of sensible heat flux (orange), latent heat flux (cyan) and net radiation at
876 the surface (green).

877

878

879 Figure 7: As in Fig. 6 for phase of the annual cycle of sensible heat flux (orange) latent heat
880 flux (cyan), and net radiation at the surface (green).

881

882

883 Figure 8: As in Fig. 6 for the magnitude (brown) and phase (purple) of the annual cycle of
884 surface albedo.

885

886 Figure 9: Distribution of coupling indices for the terrestrial (x-axis) and atmospheric (y-axis)
 887 legs for the warmest consecutive 3 months of the annual cycle for FLUXNET2015 sites (white
 888 dots; identical in each panel) and for each model as indicated. Colors of dots indicate in which
 889 quadrant that FLUXNET2015 site lies: red = both indices positive; green = terrestrial
 890 positive, atmospheric negative; blue = atmospheric positive, terrestrial negative; grey =
 891 both negative. The white circle indicates the centroid of all FLUXNET2015 stations that are
 892 in that quadrant, connected by a colored dotted line to a colored circle that is the centroid of
 893 the model. Numbers in the corners of each quadrant show the number of points in that quadrant according to the model
 894 and FLUXNET2015 data, separated by a colon, and the percentage of the FLUXNET2015 sites
 895 within that quadrant that the model placed in the same quadrant. The percentage in red at
 896 the upper right of each panel is the overall percentage of sites where model and
 897 FLUXNET2015 agree on the quadrant.
 898

