Impact of Land Model Calibration on Coupled Land–Atmosphere Prediction

JOSEPH A. SANTANELLO JR.
Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

SUJAY V. KUMAR
Science Applications International Corporation, McLean, Virginia, and Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

CHRISTA D. PETERS-LIDARD
Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

KEN HARRISON
University of Maryland, College Park, College Park, and Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

SHUJIA ZHOU
Northrop Grumman Information Systems, Chantilly, Virginia, and NASA Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 22 August 2012, in final form 10 June 2013)

ABSTRACT

Land–atmosphere (LA) interactions play a critical role in determining the diurnal evolution of both planetary boundary layer (PBL) and land surface heat and moisture budgets, as well as controlling feedbacks with clouds and precipitation that lead to the persistence of dry and wet regimes. In this study, the authors examine the impact of improved specification of land surface states, anomalies, and fluxes on coupled Weather Research and Forecasting Model (WRF) forecasts during the summers of extreme dry (2006) and wet (2007) land surface conditions in the U.S. southern Great Plains. The improved land initialization and surface flux parameterizations are obtained through calibration of the Noah land surface model using the new optimization and uncertainty estimation sub-systems in NASA’s Land Information System (LIS-OPT/LIS-UE). The impact of the calibration on the 1) spinup of the land surface used as initial conditions and 2) the simulated heat and moisture states and fluxes of the coupled WRF simulations is then assessed. In addition, the sensitivity of this approach to the period of calibration (dry, wet, or average) is investigated. Results show that the offline calibration is successful in providing improved initial conditions and land surface physics for the coupled simulations and in turn leads to systematic improvements in land–PBL fluxes and near-surface temperature and humidity forecasts. Impacts are larger during dry regimes, but calibration during either primarily wet or dry periods leads to improvements in coupled simulations due to the reduction in land surface model bias. Overall, these results provide guidance on the questions of what, how, and when to calibrate land surface models for coupled model prediction.

1. Introduction

Despite evidence of the importance of land–atmosphere (LA) interactions in weather and climate prediction (e.g., Betts 2009; Seneviratne et al. 2010), the systematic impact of land surface parameterizations on coupled mesoscale modeling has proven difficult to quantify in a robust manner. The role of the land in modulating water and energy cycling has been well-documented in terms of LA coupling strength and the support of hydrological anomalies and extremes such as flood and drought (van den Hurk et al. 2011; Koster et al. 2010),...
which includes immediate effects of the land on the temperature and humidity structure in the boundary layer, convective initiation, and mesoscale circulations (Di Giuseppe et al. 2011). In addition, the influence of soil moisture on precipitation has been under community-wide investigation in a range of studies from local (Santanello et al. 2011b) to global (Koster et al. 2004) scales. What is less understood is how specific land surface models (LSMs), parameterizations, datasets, and initialization approaches impact coupled mesoscale model predictions on diurnal time scales and how each could be improved.

One confounding factor in quantifying LSM impact on coupled prediction lies in the disparate approaches to land surface spinup and initialization of community mesoscale models. Recent advances have been made in offline land data assimilation systems (LDAS; e.g., Mitchell et al. 2004; Rodell et al. 2004), with results showing improvement in prediction of ambient weather and precipitation as a result of land initial conditions taken from an LSM spinup (Chen et al. 2007; Kumar et al. 2008; Case et al. 2008, 2011; Wen et al. 2012). It still remains, though, that a great majority of coupled prediction studies do not make use of rigorous spinup or initialization methods, thereby limiting the potential impact of the land on those simulations before coupled integration even begins.

Adding to the difficulty in assessing the land surface impact on coupled modeling is that LSM physics rely heavily on diverse parameter sets corresponding to soil, vegetation, and other land-specific conditions that are difficult to measure. The accuracies of these parameters on regional scales are strongly limited by coarse-resolution datasets and their inability to capture local-scale heterogeneity in parameters such as soil hydraulic properties. As a result, attempts have been made to calibrate parameters based on observations of land surface conditions in order to ultimately improve prediction of state variables such as soil moisture (e.g., Santanello et al. 2007; Harrison et al. 2012). Fully coupled calibration studies are extremely limited, however, because of computing requirements and the difficulty in untangling the complex interactions of land and atmospheric physics and parameters.

Despite these challenges, it is important to note that the atmospheric component of a coupled model is connected to the land solely through the turbulent surface (sensible and latent heat) fluxes calculated at each time step. As a result, all the specificity and complexity of an LSM (including its parameters and the spinup approach) are invisible to the atmospheric component of a coupled simulation. A key question can therefore be asked: what are the potential impacts of providing “optimal” surface fluxes from an LSM to a coupled model versus those generated from a default or coarser-resolution (e.g., 30–100 km) land initialization? The answer would provide insight as to the first-order influence of the land surface on ambient weather (e.g., temperature, humidity, and precipitation) and coupled LA components of a prediction system [e.g., planetary boundary layer (PBL) growth and convective initiation].

This question is addressed here by combining LSM calibration and spinup approaches to produce best estimates of land surface fluxes for coupling with the Advanced Research Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2005). The focus of these experiments will be on LSM calibration over a range of surface conditions (dry to wet) in the U.S. southern Great Plains (SGP), where the land is known to have a strong modulating impact on the atmosphere (Koster et al. 2004; Dirmeyer et al. 2006). In the process, these experiments will shed light on the following issues for improving the LSM component of coupled prediction: 1) what to calibrate, 2) how to calibrate, and 3) when to calibrate. A key aspect of this work will be to comprehensively evaluate the coupled forecasts using diagnostics that simultaneously assess the land–PBL system as a whole in terms of water and energy cycling.

Section 2 of this paper provides a brief review of recent land model calibration and spinup studies, as well as the coupling diagnostics developed to assess the land–PBL system. The model, Land Information System (LIS) optimization (LIS-OPT) and uncertainty (LIS-UE) subsystems, and experimental design are then described in section 3. Results are presented in section 4, with discussion and conclusions on the role of the land surface in coupled prediction following in section 5.

2. Background

a. LSM spinup

Because in situ and remotely sensed observations of soil temperature and moisture states or fluxes are not available at the resolution of a mesoscale model grid (horizontally or vertically), LSMs are used to produce flux and state estimates based on sound physics and constrained by forcing (based on traditional atmospheric meteorological data such as precipitation) and parameter data (based on static maps of vegetation and soil properties at high spatial resolutions). The practice of long-term spinup of offline LSMs to equilibrate soil moisture and temperature states has been in place for
some time. Rodell et al. (2005) looked at the sensitivity of equilibration of total column soil moisture to the length of the spinup simulation, which was found to vary based on climate regime (e.g., cold and dry regions tend to take longer to equilibrate than warm and moist locales) and soil type.

Spinup time has also been shown to be dependent on initial values of soil moisture, atmospheric forcing, and vegetation conditions (Yang et al. 1995; Chen and Mitchell 1999; Cosgrove et al. 2003; de Goncalves et al. 2006). Overall, LSMs use either manual or automated approaches to spinup based on reaching a predefined equilibration threshold (which can range from horizontally uniform to climatologically distributed). The particular threshold values are rather arbitrary, however, and have produced spinup times varying from a few weeks to over a decade in different studies. Also a factor is whether forcing data is available to run an offline LSM for the period leading up to the coupled simulation of interest, or whether cyclical data from a single annual cycle must be used to equilibrate the states (e.g., Cosgrove et al. 2003).

Despite the diversity of spinup approaches applied, recent case studies have revealed specific impacts and improvements in coupled models as a result of improved specification of the land initial condition. Kumar et al. (2008) found significant differences in prediction of fluxes, boundary layer structure, and temperature and humidity, and improvements in precipitation forecasts when using a multiyear spinup versus using the default WRF land surface initialization. Similarly, Case et al. (2008) showed that spun-up initial conditions led to improved sea breeze circulation and 2-m temperature forecasts from WRF over Florida and improved summertime and hourly precipitation over the southeastern United States (Case et al. 2011) over the default WRF simulations.

In a similar vein, Holt et al. (2006) demonstrated a potential impact on coupled forecasts from using high-resolution representation of soil states and fluxes, while Trier et al. (2008, 2011) also show that the initial soil moisture for a WRF forecast is significantly more important than the evolution of that soil moisture during the coupled simulation itself. Using a different combination of land surface and mesoscale models, Di Giuseppe et al. (2011) indicate that consistency in the physics and configuration between offline LSM and coupled model is paramount when choosing a source for the land initialization of a coupled model.

b. Calibration of offline and coupled LSMs

As mentioned, the physics of LSMs are highly dependent on specification of a large number of parameter values representing soil, vegetation, and other surface conditions. To simplify things, lookup tables are commonly associated to a particular soil or vegetation type that relates a number of parameters to each classification. Lookup tables are only as accurate as the available soil or vegetation information, however, and attempt to provide a representative value of each parameter for each soil or vegetation type. High-resolution maps that accurately capture the observed heterogeneity in parameter values are difficult to obtain on the scales of land surface and mesoscale models (particularly for regions outside the United States), and lookup table classes do not allow for many mixed crops or soil types. This can be a problem, particularly for soils where larger differences in soil parameters have been observed within a soil type than between types (Feddes et al. 1993; Soet and Stricker 2003; Gutmann and Small 2005; Santanello et al. 2007).

To address these limitations, numerous attempts have been made to calibrate (or optimize) LSM parameters using observations of state variables such as soil moisture and surface temperature as constraints (Gupta et al. 1999; Hess 2001; Hogue et al. 2005; Liu et al. 2003, 2004, 2005; Santanello et al. 2007; Peters-Lidard et al. 2008; Harrison et al. 2012). Such approaches improve matches of state variables to observations during the calibration period (and beyond) and, in the process, address LSM systematic biases. However, it remains difficult to derive parameter information that could be evaluated independently as most studies have focused on techniques that derive large sets of “effective” parameters. Such studies also require a great deal of computational time and are limited in assessing the broader applicability of derived parameters; therefore, little has been gained in terms of quantifying the effectiveness of calibrated parameters in improving coupled simulations.

In terms of offline LSM calibration applications, Hogue et al. (2005) investigated the transferability of large calibrated parameter sets in an offline LSM and concluded that calibration results are site specific and that models should be recalibrated for changes in seasons or over longer time intervals. This suggests that if a spinup is to be used to initialize a coupled model, the calibration performed offline needs to be tailored (e.g., domain, resolution, and LSM) specifically for the experiment of interest and potentially for the period of interest as well.

Liu et al. (2003, 2004, 2005) extended parameter estimation to a semicoupled system by examining the pathways by which limitations in the LSM physics impact both offline and 1D (single column) model simulations. Each of these studies found coupled forecasts...
to be highly sensitive to the initial soil moisture value (prescribed uniformly in their study), stressing that the land initialization for coupled models is important and that the methodology of an offline spinup with calibrated parameters shows promise in providing the most accurate initial condition consistent with the surface physics and parameterizations.

c. Evaluation of LA coupling

The initial communication between the land and atmosphere occurs on local scales, and therefore, a community effort supported by the Global Energy and Water Cycle Study (GEWEX) Global Land/Atmosphere System Study (GLASS; van den Hurk et al. 2011) has been launched to diagnose and quantify local LA coupling in models, called LoCo (van den Hurk and Blythe 2008). A thorough review of LoCo research and the related diagnostic approaches can be found in Santanello et al. (2009, 2011a,b, 2013, hereafter referred to as S09, S11a, S11b, S13).

As discussed in S11a, a full understanding and quantification of LA interactions will only come by careful examination and quantification of a series of interactions and feedbacks (i.e., links in the chain) between soil moisture (SM) and precipitation (P). These relationships depend on the sensitivities of 1) surface fluxes of sensible (Qh) and latent (Qle) heat to soil moisture, 2) PBL evolution to surface fluxes, 3) entrainment fluxes to PBL evolution, and 4) the collective feedback of the atmosphere (through the PBL) on surface fluxes (Santanello et al. 2007; van Heerwaarden et al. 2009).

LIS and LIS-WRF have served as a core test bed to develop and implement LoCo diagnostics utilizing the range of LSM and PBL scheme options available in each. For example, S09, S11a, and S13 developed a model intercomparison methodology based on the “mixing diagram” theory of Betts (1992). The power of this approach lies in its ability to exploit the covariance of 2-m potential temperature and humidity to quantify the components of the SM–P relationship, and it is based only on routine variables that can be applied to any model or observation product. As shown in S09 and S11a, how anomalies and/or errors in the surface fluxes computed by a particular model are then translated into the atmospheric water and energy cycle can then be quantified using this approach. For example, results from S13 during dry/wet extremes show that the choice of LSM is critical for dry regimes, but that both PBL and LSM are comparable influences on the coupled behavior during wet regimes. LoCo diagnostics are therefore well suited to evaluate the first-order impact of land spinup on the coupled LSM–PBL system as a whole.

3. Model and site description

a. LIS and LIS-OPT/UE

The National Aeronautics and Space Administration’s (NASA) LIS (Kumar et al. 2006; Peters-Lidard et al. 2007) consists of a suite of LSMs under the same software framework and provides a detailed representation of land surface physics and states, which can then be directly coupled to an atmospheric model. The sensitivity of land surface spinups to methods and forcing data has already been addressed under this framework (Rodell et al. 2005; Kato et al. 2007). More recently, new subsystems have been added to LIS that allow sophisticated optimization and uncertainty estimation (LIS-OPT/UE) algorithms to be applied to the LSMs to exploit further the information content from observations (Kumar et al. 2012a; Harrison et al. 2012). The algorithms (e.g., Levenberg–Marquardt (Levenberg 1944; Marquardt 1963), genetic algorithm (GA; Holland 1975), and shuffled complex evolution from the University of Arizona [Duan et al. 1993]) calibrate the model parameters to observations (e.g., satellite), thereby enabling improved model forecasts and enhancing the efficiency of data assimilation approaches (Santanello et al. 2007; Peters-Lidard et al. 2008; Kumar et al. 2012a). The uncertainty estimation subsystem also includes Bayesian approaches based on Markov chain Monte Carlo (Gilks et al. 1996) to estimate the uncertainty in model parameters given calibration datasets, which enables probabilistic prediction.

Overall, the high-performance computing infrastructure in LIS provides an advantage over previous parameter estimation studies that were limited to trial and error, manual, and lower-dimensional (i.e., smaller parameter sets) calibration approaches, as demonstrated by Kumar et al. (2012a) and Harrison et al. (2012) for offline spinup and data assimilation applications. The evaluation of offline, coupled, and LIS-OPT/UE experiments is performed using an LIS-based tool called the Land Surface Verification Toolkit (LVT; Kumar et al. 2012b). LVT provides a standardized platform for intercomparing model output (from LIS or other sources) with observations and offers a range of statistical and benchmarking approaches.

b. NU-WRF

WRF-ARW has been designated as the community model for atmospheric research and operational prediction and is ideal for high-resolution (e.g., 1–10 km) regional simulations on the order of 1–14 days. WRF-ARW has an Eulerian mass dynamical core and includes a wide array of radiation, microphysics, and PBL
options as well as two-way nesting and variational data assimilation capabilities. Recently, a NASA Unified WRF (NU-WRF; https://modelingguru.nasa.gov/community/atmospheric/nuwrf) modeling system has been developed at NASA's Goddard Space Flight Center (GSFC). Built upon the WRF-ARW model, NU-WRF incorporates LIS, the WRF/Chem enabled version of the Goddard Chemistry Aerosols Radiation Transport (GOCART; Chin et al. 2000) model, GSFC radiation and microphysics schemes, and the Goddard Satellite Data Simulation Unit (SDSU; Matsui et al. 2009) into a single modeling framework.

The LA coupling is a core component of NU-WRF and has been performed through the coupling of LIS and WRF by Kumar et al. (2008). The version of NU-WRF used here includes LIS V6.2 and WRF-ARW V3.2. The advantages of coupling LIS and WRF include the ability to spin up land surface conditions on a common grid from which to initialize the regional model, flexible and high-resolution (satellite based) soil and vegetation representation, and direct coupling of the atmospheric model to the LIS subsystems (including LIS-OPT/UE). The work of S09, S11a, and S13 has demonstrated NU-WRF as a test bed for LA interaction studies and LoCo because of its land–PBL scheme flexibility and high resolution. Hereafter, we refer to NU-WRF as the coupled prediction system that includes the LIS-WRF coupling for these experiments.

The experiments described below are run on a single 500 × 500 domain at 1-km spatial resolution (see below), and include a 5-s time step, GSFC microphysics, long-wave and shortwave radiation, and the Monin–Obukhov surface layer scheme. The North American Regional Reanalysis (NARR; Mesinger et al. 2006) data were used for atmospheric initialization and lateral boundary conditions using 3-hourly nudging, and the vertical resolution of NU-WRF was specified as 43 vertical levels, with the lowest model level ~24 m above the surface. The PBL scheme selected is the Yonsei University (YSU; Hong et al. 2006) PBL, which is based on nonlocal K theory and includes explicit treatment of entrainment and counter gradient fluxes.

The LSM employed in LIS for this study is the Noah LSM version 3.2 (Ek et al. 2003), and it is identical to the version of Noah packaged in the community version of the WRF-ARW version 3.2 release. Noah is used operationally by the National Centers for Environmental Prediction (NCEP) as the LSM for the North American Mesoscale Model (NAM) and the Global Forecasting System (GFS). As such, Noah is a well-supported, developed, and utilized LSM for both offline and coupled applications. The soil type specification in LIS is based on the State Soil Geographic (STATSGO; Miller and White 1998) database over the United States, while vegetation type is assigned based on the University of Maryland (UMD) land cover dataset (Hansen et al. 2000). The combination of Noah LSM and YSU PBL is a common selection in the WRF community and has served as the default configuration for NU-WRF test cases, where it has performed well for daytime conditions over the SGP study area.

c. 2006–07 dry/wet extremes

The SGP region has been identified as a hotspot for LA coupling in terms of the strength of interactions and impact of soil moisture anomalies on clouds and precipitation (e.g., Koster et al. 2004). Because of this and the large record of observational data from the Atmospheric and Radiation Measurement (ARM) test bed (ARM-SGP), S09, S11a, and S13 have focused WRF studies on the SGP region to develop and test the LoCo diagnostics described in section 2c. In particular, S13 looked at the extreme conditions observed during the 2006–7 period and the impact on LoCo. Low anomalies of clouds and precipitation in 2006 (October–September) were immediately followed by conditions of high cloud percentage and rainfall in 2007, with 2006 being the second-driest and 2007 the seventh-wettest year on record at the ARM-SGP central facility over the period of 1921–2008 (Dong et al. 2011). This period was followed by a relatively normal summer season in 2008, with soil moisture conditions in between that of the 2006 and 2007 extremes (as confirmed by ARM-SGP observations and offline Noah simulations).

As described in S13, ideal case studies were chosen for each regime. The 14–20 July 2006 experiment consists of a lengthy dry-down period with little synoptic disturbance in which the land was free to interact and evolve with the atmosphere on primarily local scales. The case study of 14–20 June 2007 focuses on a period with scattered precipitation every 1–2 days in portions of the ARM-SGP domain, interspersed with brief dry downs in which conditions were clear and/or cloudy and culminating in a large mesoscale convective system (MCS) traversing the domain on the final nighttime period.

d. Experimental design: Default spinups

Forcing data from phase 2 of the North American Land Data Assimilation System (NLDAS-2; Xia et al. 2012) project were used to drive the spinup simulations. Noah was run offline in LIS beginning 1 January 2003, thus producing a ~3.5–4.5-yr spinup prior to the start time of the 2006 and 2007 case studies. This is slightly longer than the recommended spinup length for similar moisture and climatic regimes (soil and precipitation) and is consistent with previous studies using this LSM,
location, and time period (S09; S11a; S13) in ensuring an equilibrated soil initial condition for the coupled simulations.

Using the resultant spun-up surface fields as initial conditions for the 2006–07 case studies, NU-WRF simulations were then performed over a single high-resolution 1-km domain centered over Oklahoma and Kansas. Figure 1 shows the upper layer (0–10 cm) soil moisture values over the ARM-SGP domain as generated by Noah spinups valid at 0000 UTC on 1 July 2006, 2007, and 2008. The advantages of using LIS over default or coarse-resolution WRF initialization approaches are evident in the ability to resolve high-resolution soil moisture (Fig. 1) based on the 1-km grid inputs of vegetation and soil properties. Soil moisture varies significantly from dry to moderate (generally <25% volumetric) in 2006 to extremely wet (near saturation) and more uniformly saturated conditions in 2007, with 2008 showing more moderate soil moisture and heterogeneity.

e. Experimental design: LIS-OPT/UE case studies

The offline calibration experiments were performed using the GA algorithm in LIS-OPT/UE and applied to a set of 29 parameters describing soil, vegetation, and general characteristics in the Noah model (Table 1). The ranges were specified for each parameter based on the range across all (vegetation or soil) classes in the Noah lookup tables for each and are consistent with those used by Hogue et al. (2005) and Kumar et al. (2012a). These ranges also provide a physical constraint to ensure the realism of individual calibrated parameter values and agreement with expectations of the model physics.

The goals of calibration are to provide the best possible surface fluxes for NU-WRF simulations. Therefore, the observations employed are measurements of surface sensible (Qh), latent (Qle), and soil (Qg) heat fluxes from the ARM-SGP network of sites over the domain, including six Energy Balance Bowen Ratio (EBBR; Qh, Qle, and Qg) and 12 Eddy Correlation (ECOR; Qh and Qle only) tower locations. The GA was applied using an objective function that minimizes the root-mean-square error (RMSE) at each site with no discrimination of flux type (i.e., Qh, Qle, and Qg flux observations are weighted equally) as follows:

\[
\text{RMSE} = \sqrt{\frac{1}{N_{Qh}} \sum_{t \in T_{Qh}} (Q_{\text{lsm},t} - Q_{\text{obs},t})^2 + \frac{1}{N_{Qle}} \sum_{t \in T_{Qle}} (Q_{\text{lsm},t} - Q_{\text{obs},t})^2 + \frac{1}{N_{Qg}} \sum_{t \in T_{Qg}} (Q_{\text{lsm},t} - Q_{\text{obs},t})^2},
\]

where \( t \) is an index variable indicating the time of observation (hourly), with \( T_{Qh}, T_{Qle}, \) and \( T_{Qg} \) indicating the sets of observation times for each observation type and \( N_{Qh}, N_{Qle}, \) and \( N_{Qg} \) indicating the number of observations of each type. The subscripts denote the surface fluxes from the land surface model (lsm) and observations (obs). The calibration was performed over the periods 1 May to 1 September of 2006, 2007, and 2008 to produce separate calibrated parameter sets for the dry, wet, and normal regimes. Having three separate calibration periods allows for the study of the impact of calibration period and varying atmospheric and land surface conditions on the calibration results.

The number of observations of Qle, Qh, and Qg that are used in the GA optimization are comparable, but vary slightly from 2006 (\( N_{Qle}, 48,546; N_{Qh}, 48,822; N_{Qg}, 32,218 \)) to 2007 (\( N_{Qle}, 37,936; N_{Qh}, 39,063; N_{Qg}, 30,100 \)) and to 2008 (\( N_{Qle}, 45,767; N_{Qh}, 48,353; N_{Qg}, 31,344 \)). As a result, the objective function is skewed toward the fluxes with the greater number of observations in each case and is weighted more heavily toward Qh and Qle than Qg, which is available only from EBBR stations. It should also be noted that EBBR estimates both Qh and Qle simultaneously through use of the Bowen ratio method, and although they offer a desirable constraint on the calibrated flux values, they do not offer truly independent sources of information to the objective function.

The GA integrations use a population size of 50 and employ an elitism strategy to ensure that the current best solution is not overwritten during GA evolution, with a mutation rate of 0.005 and a recombination rate of 0.9. The GA parameters (including the mutation and recombination rates) are chosen largely from experience and the success of the optimization simulations in Kumar et al. (2012a). The algorithm was found to converge after approximately 200 generations, when the fitness of the best solution was found not to improve in the last 30 generations.

From these simulations, a unique calibrated value of each of the 29 Noah parameters was obtained at each of the 18 grid cells pertaining to the flux sites. To obtain calibrated values covering the full model domain, the values from each site were grouped and averaged by common vegetation and soil types and assigned to the full domain based on the vegetation and soil classification.
at each grid cell. Note that Noah parameters were designated into soil (15 parameters, five classes in the SGP domain), vegetation (11 parameters, three classes in the SGP domain), and general (three parameters, no classification) categories as based on their functionality and most direct impact on the model physics. For example, for a soil-related variable such as porosity, the calibrated values of porosity from each flux site with a “clay” classification were averaged and then applied as the porosity value to the remainder of the domain where clay was also the soil type. Also, if a soil/vegetation class occurs in the domain but was not represented at one of the observation sites, default table values are used. General parameters are constant across the domain and do not have a classification, and therefore were averaged across all the sites.

Using the calibrated parameters, new soil, vegetation, and general lookup tables for Noah were then generated. Spinup runs (as described in the previous section for the default case) were repeated using the new tables based on the 2006, 2007, and 2008 calibration results, thereby producing spun-up and initial conditions that are optimized for dry, wet, and average conditions, respectively, over this region. To examine the impact of calibrated spinups on coupled forecasts, four targeted NU-WRF case studies were then chosen from the larger 7-day periods described above, with characteristics as follows:

1) 14 July 2006: 24 h, dry regime (NU-WRF test case);
2) 18–19 July 2006: 48 h, dry regime (end of dry-down period);
3) 16–17 June 2007: 48 h, wet regime (limited/scattered precipitation);
4) 19–20 June 2007: 48 h, wet regime (scattered/MCS precipitation).

Each simulation was initialized at 1800 LST with the land using the LIS soil moisture and temperature states from each calibration run (described below) and the atmosphere using the NARR analysis as generated from the WRF Preprocessing System (WPS).

NU-WRF was then run for each case study above using four different combinations of parameter values/lookup tables, as shown in Table 2. The array of simulations was designed to capture the impact on NU-WRF forecasts from using a combination of 1) default spinup (uncalibrated) and default parameters in the coupled run (DEF), 2) default spinup with calibrated parameters in the coupled run (CPL), 3) calibrated spinup with default parameters in the coupled run (SPN), and 4) calibrated spinup with calibrated parameters in the coupled run (SCP). Note that the focus of the results presented here will be on the differences between the DEF (no
calibration) and SCP (fully calibrated) cases, but CPL and SPN offer the ability to parse out the relative impacts of using optimal parameters during the spinup versus coupled simulation period and will be included in the discussion when relevant.

f. Observation data

The ARM-SGP program provides a long-standing record of quality-controlled surface flux, meteorological, and hydrological observations along with atmospheric profiles for a network of sites across the domain shown in Fig. 1. This includes collocated soil moisture; net radiation; and sensible, latent, and soil heat, along with collocated surface meteorology data. For the calibration experiments, ARM-SGP data were collected from ECOR and EBBR towers as described above. Typical error ranges for Qle and Qh are $\sim 10\%$ for EBBR with perfect closure (by definition) and about 5%-6% for ECOR with 75%-90% closure (http://www.arm.gov/instruments/ecor; Wilson et al. 2002). The implications of observational bias on the results will be discussed in section 5.

The LoCo evaluation was performed using collocated surface meteorology, flux towers, and radiosonde profile data. For the mixing diagram analysis, we have defined the residual vector in the diagrams (formerly the “entrainment vector” as in S09) as the atmospheric response vector ($\mathbf{V}_{\text{atm}}$) to more precisely reflect the inherent assumptions. The vector $\mathbf{V}_{\text{atm}}$ represents the full sum of atmospheric contributions to LA coupling in terms of the PBL budget fluxes. A significant fraction of $\mathbf{V}_{\text{atm}}$ is composed of entrainment fluxes, but it also incorporates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMCMAX</td>
<td>Porosity (-)</td>
<td>0.30</td>
<td>0.50</td>
</tr>
<tr>
<td>PSISAT</td>
<td>Saturated matric potential (-)</td>
<td>0.01</td>
<td>0.70</td>
</tr>
<tr>
<td>DKSAT</td>
<td>Saturated hydraulic conductivity ($\text{m s}^{-1}$)</td>
<td>$5.71 \times 10^{-6}$</td>
<td>$2.33 \times 10^{-5}$</td>
</tr>
<tr>
<td>DWSAT</td>
<td>Saturated soil diffusivity (-)</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>BEXP</td>
<td>Beta parameter (-)</td>
<td>40</td>
<td>1000</td>
</tr>
<tr>
<td>QUARTZ</td>
<td>Soil quartz content (-)</td>
<td>0.10</td>
<td>0.90</td>
</tr>
<tr>
<td>RSMIN</td>
<td>Minimum stomatal resistance (m)</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>RGL</td>
<td>Parameter used in solar radiation term of canopy resistance (-)</td>
<td>0.05</td>
<td>6.00</td>
</tr>
<tr>
<td>HS</td>
<td>Parameter used in vapor pressure deficit term of canopy resistance (-)</td>
<td>0.10</td>
<td>2.00</td>
</tr>
<tr>
<td>Z0</td>
<td>Roughness length (m)</td>
<td>0.10</td>
<td>2.00</td>
</tr>
<tr>
<td>LAI</td>
<td>Leaf area index (-)</td>
<td>0.05</td>
<td>6.00</td>
</tr>
<tr>
<td>CFACCTR</td>
<td>Canopy water parameter</td>
<td>0.10</td>
<td>2.00</td>
</tr>
<tr>
<td>CMCMAX</td>
<td>Maximum canopy water content (m)</td>
<td>$2.00 \times 10^{-5}$</td>
<td>$2.00 \times 10^{-5}$</td>
</tr>
<tr>
<td>SBETAT</td>
<td>Parameter used in the computation of vegetation effect on soil heat flux (-)</td>
<td>$1.00 \times 10^{-4}$</td>
<td>$1.00 \times 10^{-4}$</td>
</tr>
<tr>
<td>RSMAX</td>
<td>Maximum stomatal resistance (m)</td>
<td>5000</td>
<td>10000</td>
</tr>
<tr>
<td>TOPT</td>
<td>Optimum transpiration air temperature (K)</td>
<td>293</td>
<td>303</td>
</tr>
<tr>
<td>REFDK</td>
<td>Reference value for saturated hydraulic conductivity ($\text{m s}^{-1}$)</td>
<td>$5.00 \times 10^{-7}$</td>
<td>$3.00 \times 10^{-5}$</td>
</tr>
<tr>
<td>FXEXP</td>
<td>Bare soil evaporation exponent (-)</td>
<td>0.20</td>
<td>4.00</td>
</tr>
<tr>
<td>REFDT</td>
<td>Reference value for surface infiltration (-)</td>
<td>0.10</td>
<td>10.00</td>
</tr>
<tr>
<td>CZIL</td>
<td>Parameter used in the calculation of roughness length of heat (-)</td>
<td>0.05</td>
<td>0.80</td>
</tr>
<tr>
<td>FRZK</td>
<td>Ice threshold (-)</td>
<td>0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>SNUP</td>
<td>Snow depth threshold (m)</td>
<td>0.025</td>
<td>0.08</td>
</tr>
<tr>
<td>SMCMREF</td>
<td>Reference soil moisture where transpiration stress begins (-)</td>
<td>0.00</td>
<td>0.50</td>
</tr>
<tr>
<td>SMCDRY</td>
<td>Dry soil moisture threshold where direct evaporation from top layer ends (-)</td>
<td>0.00</td>
<td>0.15</td>
</tr>
<tr>
<td>SMCWLT</td>
<td>Wilting point (-)</td>
<td>0.00</td>
<td>0.15</td>
</tr>
<tr>
<td>F1</td>
<td>Soil thermal diffusivity coefficient (-)</td>
<td>$1.26 \times 10^{6}$</td>
<td>$3.56 \times 10^{6}$</td>
</tr>
<tr>
<td>CSOIL</td>
<td>Soil heat capacity for mineral soil component (-)</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>SLOPE</td>
<td>Linear reservoir coefficient (-)</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>EMISS</td>
<td>Surface emissivity (-)</td>
<td>0.80</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 2. Description of calibration approaches and parameter sets used in NU-WRF simulations.

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Description</th>
<th>Spinup parameters</th>
<th>Coupled parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DEF</td>
<td>Default run with uncalibrated parameters in LIS and NU-WRF</td>
<td>Default</td>
</tr>
<tr>
<td>2</td>
<td>CPL</td>
<td>Impact of calibrated parameters in NU-WRF only</td>
<td>Default</td>
</tr>
<tr>
<td>3</td>
<td>SPN</td>
<td>Impact of calibrating LIS spinup (ICs) only</td>
<td>Calibrated</td>
</tr>
<tr>
<td>4</td>
<td>SCP</td>
<td>Impact of full calibration (LIS and NU-WRF)</td>
<td>Calibrated</td>
</tr>
</tbody>
</table>
horizontal advection and reflects that the use of 2-m temperature and humidity (as opposed to PBL mean quantities) in the mixing diagrams can proportionally overestimate the magnitude of the residual vector component fluxes. While approaching the residual vector in this manner prohibits the ability to quantify entrainment fluxes in absolute terms (or in comparison to prior analyses or observations of the entrainment ratio, for example), the mixing approach still allows the coupled system (including the bulk PBL response) to be evaluated consistently across model runs and observations.

4. Results

The performance of the offline calibration experiments will be evaluated first, followed by the impact of
spinup calibration and initialization on NU-WRF predictions and LoCo and the sensitivity of the coupled results to the period of calibration, and concluding with the uncertainty introduced into the forecasts by different parameter sets.

a. Offline calibration

Before examining the coupled cases, it is important to quantify the impact of the calibrated parameters on the offline spinup. Figure 2 shows the flux components simulated using default and calibrated Noah parameters during the dry regime (2006) versus observations at each of the ARM-SGP sites and over the full domain. Both Qh and Qle show improvement at nearly all sites, with RMSE values reduced by up to 25.7 W m$^{-2}$ (10.5 W m$^{-2}$ on average) in Qle, and up to 45.3 W m$^{-2}$ (19.1 W m$^{-2}$ on average) in Qh. Note that the 95% confidence interval for the average error across all sites is about 4–7 W m$^{-2}$, so the improvements are statistically significant. The improvement due to the calibration is also clearly evident in the mean diurnal cycle behavior of Qh and Qle across all sites. Focusing on the daytime when the turbulent fluxes are large and positive, Qh matches observations almost exactly and improves over the high
bias present in the default simulations. Analogously, daytime Qle increases because of calibration and matches observations more closely than when default parameters are used in Noah. The Noah model has often been shown to produce systematic over/underestimation of surface fluxes, and the GA calibration successfully improves upon the biases exhibited for the SGP and study period demonstrated here.

The surface soil heat flux Qg shows more mixed results, with five of the 11 EBBR sites showing slight degradation after calibration, but the magnitudes for Qg are small overall, and this does not present a concern for this study. The mixed results are partially a reflection of the reduced number of observations of Qg available for the GA and the heavier weighting toward Qh and Qle. In addition, phase errors in Qg are well documented (Robock et al. 2003; Reichle et al. 2010) and could possibly be corrected if joint calibration approaches including soil temperature and Qg were conducted.

Figure 3 shows the offline calibration results for the wet regime (2007), and once again, Qh and Qle are improved at nearly all ARM-SGP sites (and in the case of Qh, all sites show improvement). In this calibration, Qh improvements are more modest than in 2006 (up to 25.9 W m\(^{-2}\) and 12.3 W m\(^{-2}\) on average), while Qle improvements are larger than during the dry regime (up to 54.9 W m\(^{-2}\) and 12.3 W m\(^{-2}\) on average). Interestingly, site E24 shows the largest improvement in this case, opposite of the 2006 calibration. The mean diurnal cycles show marked improvement (decrease) in daytime Qle over the default simulations, while Qh is only very slightly impacted (and also decreased). This suggests an available energy bias in the NLDAS-2 forcing data and subsequent overestimation in the offline Noah runs in 2007. Once again, Qg shows mixed results as five of 11 sites show degradation, though in this case there is a noticeable increase in Qg after calibration that improves afternoon simulations but does not impact the phase error where Qg peaks too early (as in the 2006 case).

Overall, the largest impact and improvement due to calibration of Noah is seen in Qh in 2006 and in Qle in 2007. Physically, this can be explained by the fact that during the dry regime, Noah has a dry bias and produces too little evaporation, thereby overestimating Qh. In the wet regime, Noah has a wet bias and produces too much Qle (partially due to too much net radiation). These results are also consistent in that, during a dry regime, which is water-limited, the primary adjustment in fluxes would be toward the higher magnitude flux (Qh), and during a saturated regime, the largest impact would be felt in Qle. In fact, the domain-averaged statistics show a maximum reduction (relative to the other flux components) in positive bias of 26.46 W m\(^{-2}\) in Qh during 2006 and 31.4 W m\(^{-2}\) in Qle during 2007. We therefore infer that the calibration has adjusted the parameter set to correct for the dry bias in 2006 by modifying the efficiency of the evaporative physics in Noah (and vice
versa in 2007) to complement the new soil moisture levels and produce the optimal fluxes based on observations.

b. Coupled simulations

To assess the impact of offline LSM calibration on the coupled system, LoCo diagnostics are used to simultaneously evaluate the land (LSM) and atmospheric (PBL) component evolution and interaction.

1) 14 JULY 2006

The mixing diagram analysis for the 14 July 2006 case at the ARM-SGP E4 site is shown in Fig. 4. Focusing first on the comparison of the DEF and SCP simulations, it is shown that the default Noah parameters produce the poorest simulation of heat and moisture states and fluxes in NU-WRF. Visually, the DEF curve is drier (and slightly warmer) than observed throughout the daytime period. This is improved significantly in the SCP simulation, which matches closely with observed 2-m potential temperature (T2) and 2-m specific humidity (Q2) throughout. Table 3 provides error statistics of simulated versus observed T2 and Q2 coevolution, and because mixing diagrams are in energy space, these can be represented in units of J kg\(^{-1}\) and can be used to describe a total RMSE,

\[
\text{Total RMSE} = \sqrt{\frac{1}{T} \sum_{i=1}^{T} (T_{wrf,i} - T_{obs,i})^2} + \sqrt{\frac{1}{T} \sum_{i=1}^{T} (Q_{wrf,i} - Q_{obs,i})^2}, \tag{2}
\]

and a total mean absolute error (MAE),

\[
\text{Total MAE} = \frac{1}{T} \sum_{i=1}^{T} |T_{wrf,i} - T_{obs,i}| + \frac{1}{T} \sum_{i=1}^{T} |Q_{wrf,i} - Q_{obs,i}|, \tag{3}
\]

![Fig. 5. Daytime mean evaporative fraction vs PBL height for the simulations in Fig. 4.](image-url)
of heat and moisture combined (i.e., quantifying the spatial differences between the model and observed curves in Fig. 4), where the subscripts wrf and obs indicate the NU-WRF–simulated and ARM-SGP–observed values of each and \( t \) is the number of hours evaluated in Fig. 4. These metrics confirm that the DEF run performs worst of all the simulations, while the SCP improves all aspects of the temperature and moisture states (T2 and Q2) by 15%–26% in total RMSE and 8%–30% in bias.

The fluxes in the coupled system can be evaluated in Fig. 4 via the surface (\( \mathbf{V}_{\text{sfc}} \)) and residual (\( \mathbf{V}_{\text{atm}} \)) vectors and their flux components and ratios. Note that fluxes into the PBL are defined as positive, as sources of energy for the PBL. As expected, SCP produces a surface Bowen ratio (\( \beta_{\text{sfc}} = \frac{Q_h}{Q_{le}} \)) nearly identical to that observed because of the calibration to surface fluxes performed, which produced the parameters used in the SCP simulation. DEF overestimates \( \beta_{\text{sfc}} \), consistent with the dry bias observed in the offline spinup and the coupled T2 and Q2 results. The components of \( \mathbf{V}_{\text{atm}} \) are also impacted by the LSM calibration by \( \sim 15\% \) and are slightly closer to observations. Likewise, the relative proportions of atmosphere-to-land fluxes (defined as the ratios of atmospheric fluxes to surface fluxes generated; \( A_{le} = \frac{Q_{le_{\text{atm}}}}{Q_{le}} \) and \( A_h = \frac{Q_{h_{\text{atm}}}}{Q_h} \)) show substantial improvement in SCP over default, where the higher Qle

![Fig. 6. Diurnal cycle and error statistics of T2 and Q2 for the 14 July 2006 case and evaluated at 214 station pairs across the ARM-SGP domain for each 6-hourly increment.](image1)

![Fig. 7. Qle, Qh, and Qg RMSE and bias statistics (W m\(^{-2}\)) on 14 July 2006 measured against the full set of ARM-SGP flux tower sites and evaluated hourly.](image2)
and lower Qh produce better ratios of land to PBL fluxes as a result of correcting the dry bias at the surface. [It should be noted here that advection can be incorporated explicitly into mixing diagram analyses as described in S09, but for the case studies evaluated here, the advection ratios (as defined in S11a) were sufficiently small such that they are only a very small proportion of $V_{atm}$.]
Focusing on the remaining two simulations, CPL and SPN, indicates how calibrated parameters impact coupled simulations when used in either offline spinups or the coupled run only. It is first evident that SPN does well with T2 and Q2 state estimation, correcting the dry bias of Noah and producing the best overall error metrics in Table 3. The fluxes of SPN are severely overcompensated, however (e.g., \( \beta_{de} \) very low), and produce too much evaporation. Because the calibrated parameters in this simulation are used only for the spinup, these results indicate that the default parameters still employed in the coupled run produce too high evaporation rates for the given initial soil moisture state. The CPL simulation performs poorly both in terms of T2 and Q2 (with comparable or worse metrics in Table 3 to the DEF simulation) and surface and PBL fluxes, indicating that using calibrated parameters only for the coupled simulation along with a default spinup does not impact or improve the coupled forecast at all. These results are also consistent with those of Trier et al. (2008), who showed that initial soil moisture (i.e., fluxes calibrated in SPN) has a much larger influence on forecasts than the evolution of soil moisture during the coupled run (i.e., fluxes calibrated in CPL).

A related diagnostic of the coupled system performance is the relationship of daily midday evaporative fraction (EF) and daytime maximum PBL height (PBLH), as shown in Fig. 5. Once again, the best combination of land and atmospheric behavior is exhibited by the SCP simulation, which closely matches both the EF (which integrates the land surface condition) and PBLH (which integrates the atmospheric response). SPN and CPL are the extremes in terms of EF and PBLH, while the dry bias in the DEF simulation is evident and leads to slightly higher PBL growth.

From the full suite of simulations and diagnostics in Figs. 4 and 5 and Table 3, it is clear that offline LSM calibration can improve coupled simulation components significantly and in a consistent fashion in terms of correcting a bias and the impact of that correction (e.g., soil moisture) on the coupled components (e.g., T2 and Q2). It is also evident that employing calibrated parameters in both the offline spinup and the coupled run is required to achieve optimal improvement in coupled prediction. It is the combination of a spinup produced with calibrated parameters that supports a wetter initial condition and those same parameters that support lower evaporation rates in the coupled simulation that are compensatory. Therefore, if the calibrated parameters are only used in either the spinup or coupled run, significant and overreaching impacts will be seen in the prediction of coupled states and/or fluxes (as seen in SPN and CPL).

### Table 4. Error statistics from (a) Fig. 9a and (b) Fig. 9b for all four simulations. Bold values as in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>DEF</th>
<th>CPL</th>
<th>SPN</th>
<th>SCP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total RMSE</td>
<td>6018.59</td>
<td>5992.34</td>
<td>3977.58</td>
<td><strong>5086.32</strong></td>
</tr>
<tr>
<td>Total MAE</td>
<td>4921.32</td>
<td>4992.19</td>
<td>3050.16</td>
<td><strong>4129.53</strong></td>
</tr>
<tr>
<td>Bias Q2</td>
<td>−7889.19</td>
<td>−7859.74</td>
<td>−5002.86</td>
<td><strong>−6663.78</strong></td>
</tr>
<tr>
<td>Bias T2</td>
<td>1953.45</td>
<td>2124.63</td>
<td>818.18</td>
<td><strong>1595.27</strong></td>
</tr>
<tr>
<td>N-S efficiency</td>
<td>−0.385</td>
<td>−0.373</td>
<td>0.394</td>
<td><strong>0.011</strong></td>
</tr>
</tbody>
</table>

A robust measure of the impact of LSM spinup and calibration on weather prediction can be found in the performance of T2 and Q2 across the entire model domain. Figure 6 shows the domain-averaged statistics computed using the Model Evaluation Tools (MET) statistical software package [developed by the National Center for Atmospheric Research (NCAR; www.dtcenter.org/met/users/docs/overview.php) and incorporating NCEP Automated Data Processing (ADP) atmospheric and surface data] and based on 214 site observations at 6-hourly intervals on 14 July 2006, which provides a true independent evaluation of the model. In particular, the total RMSE and bias statistics are largely improved in SCP versus DEF and are consistent in terms of lowering the dry/warm bias of the default simulation. Also plotted are the results from a NU-WRF simulation that does not use LIS or a spinup of the Noah LSM (as a true “off the shelf” WRF default case comparison).

Overall, by introducing a spinup (DEF versus WRF), there is a definite increment of forecast improvement over using the default and coarser atmospheric initialization data source (e.g., NARR in this case). Performing offline calibration for a spinup then increases the accuracy of the simulation even further (SCP versus DEF versus WRF). Likewise, the land surface energy balance (Qh, Qle, and Qg) components across the entire suite of 19 ARM-SGP sites are shown in Fig. 7, where improvement is seen across the board in terms of reducing the total RMSE and bias. These results provide strong evidence that spinup and calibration improves coupled forecasts across the entire NU-WRF domain, as well as the individual site details shown in Figs. 4 and 5.

2) 18–19 JULY 2006

The other dry regime case study results are shown in Fig. 8 and Table 4. As the dry down has progressed over the period, there is a larger diurnal range in T2 observed...
(~20 K) than the 14 July case (~13 K), while the humidity ranges are comparable on 18 July but reach a much drier condition on 19 July as the surface nears desiccation. On both days in Fig. 8, the DEF simulation shows a more extreme dry bias now versus observations, as reflected in Q2 and the surface Bowen ratio. Despite this, the calibration in SCP still produces consistent improvement in heat and moisture states and fluxes,

Fig. 9. As in Fig. 8, but for (top) 16 and (bottom) 17 June 2007.
particularly on 18 July. The value of $\beta_{sfc}$ on 19 July is observed to be much higher than the previous day and supports a sharp diurnal decrease in Q2 due to lack of surface evaporation and increased dry air entrainment (and is similar to the mixing diagram signature seen in the dry soils results of S09 and S11a). Overall, the SPN simulation (not shown) produces the lowest T2 and Q2 errors, but as was the case for 14 July, this occurs for the wrong reasons, as $\beta_{sfc}$ is vastly underestimated while CPL remains close to the DEF results.

That SCP does not match or improve $\beta_{sfc}$ observations as well as the previous cases is because the overall nature of the calibration is to correct the dry bias in Noah, thereby increasing the soil moisture and Qle. The calibration works well overall, but for extreme conditions like on 19 July, the DEF simulation just so happens to produce better $\beta_{sfc}$ because of its inherent dry bias. The limits of calibrating the spinup are also evident here, as the shift due to higher initial soil moisture is felt in the coupled simulation to the degree of the shift in DEF to SCP curves and suggests that there remain deficiencies in LSM physics that limit the quality of results even after a detailed calibration is performed.

3) 16–17 JUNE 2007

The wet regime cases show a vastly different signature in the mixing diagrams that reflects much higher evaporation rates at the surface and limited PBL growth and entrainment above. Figure 9 and Table 5 show that the DEF simulations generally perform well relative to observations in terms of T2 and Q2 evolution. As a result, there is very little impact of using calibrated versus default parameters, though the patterns are consistent in that CPL performs worst and SPN performs best in terms of T2 and Q2 metrics. The calibration does improve $\beta_{sfc}$ in SCP over DEF and is very close to observations, as should be the case based on the calibration design (note that the calibration performed for these cases was appropriately based on the 1 May to 1 September 2007 period). There is not any translation of this improvement to the PBL fluxes or 2-m states, however. This is consistent with the results of S13, who showed that the impact of a particular LSM is dampened during wet regimes when the PBL scheme and atmosphere-dominated regime take over. Furthermore, when the LSM and coupled model perform well [as 16 June total RMSE, total MAE, bias, and Nash–Sutcliffe (N-S) metric suggest], there is little to be gained in calibrating large sets of parameters because the inherent predictability in the system has already been maximized.

4) 19–20 JUNE 2007

At the end of the wet regime, much poorer performance is seen in both the DEF and SCP simulations (Fig. 10, Table 6) in terms of the diurnal evolution of T2 and Q2. Particularly on 19 June, when DEF has a wet bias in the morning, there is degradation across all metrics (with the exception of the Q2 bias), which is again consistent with the calibration attempt to correct the overall dry bias that is not evident on this particular day. As also evident from the comparisons of all the case studies thus far, there is a noticeable shift on 19 June to a very wet regime (high Q2) that reflects frequent precipitation events in the days prior (including the passage of a MCS over the study region).

The performance on 20 June is similar to 16–17 June in that there is very little impact of calibration on the results. Overall, the wet regime is dominated by low $\beta_{sfc}$ and relatively high Qle, along with lower net radiation (due to clouds and precipitation) and reduced PBLH, entrainment, and diurnal cycles of T2 and Q2. This makes the potential impact from LSM adjustments (such as calibration, spinup, and initialization approaches) on the coupled system much lower than in the dry regime. In addition, the attempt of calibration to systematically reduce inherent LSM biases works least well for the extremes of regimes (e.g., just after frequent rainfall or at the end of a severe dry down) as opposed to the more benign, moderate, and transitional periods (as reflected in the overall offline and domain-averaged results presented above).

c. Period of calibration

The second part of this analysis addresses the question, what is the impact of the period of calibration on coupled predictions? The 2006 case studies above were performed using parameters calibrated during the summer 2006 period, and the 2007 cases were performed with parameters calibrated during 2007. For broader applicability of this methodology, it is important to address the
impact of data availability and limitations on the calibration. For example, if observed fluxes are only available for a limited time or for a certain year or season (as is often the case for field experiments) that does not coincide with the forecast period of interest, there likely will not be as optimal results seen in the offline calibration or coupled simulations. In addition, application of parameters calibrated outside the forecast period of

![Graph](image)

Fig. 10. As in Fig. 8, but for (top) 19 and (bottom) 20 June 2007.
improvement in both Qle and Qh, nearly matching

5 C678 Impact of calibrating to all three years combined 2006, 2007, 2008

3 C07 Impact of calibrating during 2007 only 2007

2 C06 Impact of calibrating during 2006 only 2006

1 DEF Default run with uncalibrated parameters Default

C07 is the same as SCP in Figs. 9 and 10, and DEF is the
parameters in the spinup and during the coupled run,
and therefore, C06 is identical to SCP in Figs. 4 and 8,
These simulations are each conducted using calibrated

The results for the offline calibration using all three
simulations with different year calibrations are shown in
Figs. 13 and 14 and Table 8. DEF and C06 are the same
as in Fig. 4, but what is now evident is the spread in re-
results introduced by different calibration periods. C07
performs nearly as well as C06 despite that this is a 2006
case (Fig. 14), with both the T2 and Q2 evolution and
error metrics almost identical (Table 8). The similarity
of C06 and C07 follow in the EF versus PBLH analysis
(Fig. 14) as well. The worst-performing simulation by far
is that with the calibrated parameters from the average
year (C08), which is too dry and significantly overesti-
mates βfic as a result (low Qle, high Qh). This translates
into an atmospheric response that is too large and is
reflected in low EF and large PBL growth in Fig. 14. The
calibration using all three years of data (C678) generally
performs well, but less so than either C06 or C07, which
is as expected given the performance and weighting of
the individual years.

These results show that impacts on coupled prediction
during extremes are maximized when the calibration is
able to correct the LSM bias most significantly. In this
case, C06 and C07 occur during primarily dry and wet
extremes when the offline LSM produces large errors in
fluxes, and therefore, the resultant calibrated param-
eters provide more improvement in the coupled predic-
tion during similar periods. This will be discussed further
in section 5. On the contrary, calibration during an
overall average year such as C08 (even one that has
short precipitation and dry downs within it) when the
LSM performs well does not yield much improvement
when applied to a dry extreme case. Similar results are
also seen for the 18–19 July 2006 case study (ranked as
C06, C07, C678, and C08 from most to least improve-
ment), and similar mixed/limited impacts are seen in the
2007 cases because of the atmospherically controlled
conditions. It should be investigated further whether this
“targeted” calibration approach works for other loca-
tions and conditions as well, but it is suggested that
a single long-term calibration period (multiyear) might

interest can be evaluated here as truly independent
samples.

Table 7 lists the experiments conducted to determine
the impact of having observations only during dry, wet,
or average years, or having all three years available.
These simulations are each conducted using calibrated
parameters in the spinup and during the coupled run,
and therefore, C06 is identical to SCP in Figs. 4 and 8,
C07 is the same as SCP in Figs. 9 and 10, and DEF is the
same as in all previous analyses.

The land surface energy balance components for the
2008 offline calibration are shown in Fig. 11. Im-
provement in total RMSE of Qle and Qh is seen at all but
day three and five sites, respectively, but to a much lesser
degree overall (3–10 W m−2) than was seen in 2006
and 2007. Likewise, the impact of calibration on the
diurnal cycle fluxes is very small, particularly for Qle
(which is already simulated quite well by default), al-
though Qg shows more impact and degradation during
daylight than either 2006 or 2007. Likewise, the impact of calibration on the

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Description</th>
<th>Spinup parameters</th>
<th>Coupled parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DEF</td>
<td>Default</td>
<td>Default</td>
</tr>
<tr>
<td>2</td>
<td>C06</td>
<td>Impact of calibrating during 2006 only</td>
<td>2006</td>
</tr>
<tr>
<td>3</td>
<td>C07</td>
<td>Impact of calibrating during 2007 only</td>
<td>2007</td>
</tr>
<tr>
<td>4</td>
<td>C08</td>
<td>Impact of calibrating during 2008 only</td>
<td>2008</td>
</tr>
</tbody>
</table>

TABLE 7. Description of calibration approaches and parameter sets used in NU-WRF simulations.

<table>
<thead>
<tr>
<th>Total RMSE</th>
<th>3501.51</th>
<th>3483.16</th>
<th>3576.48</th>
<th>3987.41</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias T2</td>
<td>−257.51</td>
<td>−1412.37</td>
<td>1159.99</td>
<td>142.81</td>
</tr>
<tr>
<td>Bias Q2</td>
<td>3213.09</td>
<td>2043.18</td>
<td>2811.18</td>
<td></td>
</tr>
<tr>
<td>N-S efficiency</td>
<td>−1.193</td>
<td>−2.096</td>
<td>−1.285</td>
<td>−1.673</td>
</tr>
</tbody>
</table>

**TABLE 6. Error statistics from (a) Fig. 11a and (b) Fig. 11b for all four simulations. Bold values as in Table 3.**
not generate parameters that are suitable to extreme period prediction.

d. Uncertainty propagation

An interesting question that is inherent in parameter estimation studies is how to quantify the sensitivity of LSMs to calibrated parameter sets generated by algorithms such as GA. To address this issue, an additional suite of simulations was conducted using a simple Monte Carlo simulation (MC-SIM) sampling algorithm implemented in LIS-OPT/UE in order to propagate uncertainty from inputs (e.g., soil, vegetation, and general parameters) to model outputs (e.g., offline spinup and coupled prediction). As such, this algorithm allows for an assessment of LSM uncertainty and can be used to gauge the relative sensitivity of the coupled system to LSM inputs. A small sample size (five) was applied, given that WRF does not have a true ensemble mode and essentially requires independent integrations for each set. As in Kumar et al. (2012a), uniform distributions were applied to all parameters given the physically realistic ranges in Table 1 (also based on Kumar et al. 2012a). The result is a sense of the spread in simulations prior to calibration.
Figure 15 shows the results of the DEF and C06 simulations (as in Fig. 14) for the 14 July 2006 case, along with the simulations using the five parameter sets sampled with MC-SIM (used in both the spinup and coupled run, as for C06). The large spread in results (shaded area) highlights the importance of LSM parameter sets in the coupled forecast of heat and moisture states and fluxes. That MC-SIM randomly sampled these sets suggests the full spread, using physically reasonable bounds on parameter values as was done here, could actually be much larger than shown here as well. Nearly all of the MC-SIM simulations are on the dry side of observations, an indication of the dry bias in the Noah model that is only circumvented when using the full C06 calibration with observations. The fluxes in MC-SIM vary quite a bit, where $\beta_{sf}$ ranges from 0.733 to 4.960 and large errors versus observed are carried into the atmospheric components of the system.

5. Discussion

The questions addressed in this study of improving coupled prediction using LSM calibration have shed light on the following issues: 1) what to calibrate, 2) how
to calibrate, and 3) when to calibrate. Because fluxes are the most important aspect of LSMs for atmospheric models, the largest impact will be seen in calibrating an LSM to Qle and Qh observations. In the approach presented here, in contrast to Santanello et al. (2007), we calibrate only fluxes, and therefore, soil states such as moisture and temperature are by-products without observational constraints. Current and future satellite missions will provide soil moisture state observations that can be used to calibrate soil hydraulic properties as shown in Santanello et al. (2007). It is therefore increasingly likely that multi-objective calibration approaches (e.g., Gupta et al. 1999) will be most beneficial to LA prediction if both state and flux measurements can be used simultaneously in LSM calibrations.

With regards to the calibration of fluxes themselves, it is also important to acknowledge the inherent uncertainty in measurements by current best-available instruments such as EBBR and ECOR. As described in section 3, there is up to 10% error (or larger) in Qh and Qle, as well as up to 25% closure gaps possible in these data. Therefore, we have performed preliminary analyses (not shown) of the potential impact of this uncertainty on the evaluation of the coupled simulations and mixing diagrams. Results show that ±10% changes in the surface flux vector components only lead to marginal impacts on atmospheric response and derived ratios. More importantly, this uncertainty does not change any of the conclusions that the full calibration (SCP) simulation produces results closest to the observed range of states and fluxes relative to the DEF or other mixed calibration simulations.

In terms of how to calibrate, it is not the algorithm choice (e.g., similar performance has been seen in LIS-OPT/UE intercomparisons of the three methods therein; Harrison et al. 2012) so much as the parameter sets and mapping approach that are employed that is important for coupled prediction. NU-WRF is fully 3D and communicates horizontally between grid cells through the atmospheric flow. This is in contrast to LIS and most LSMs, which operate in one dimension. This makes it particularly important that parameter calibration and assignment be considered carefully for coupled studies. The approach performed in this study entailed the assignment of soil, vegetation, and general parameter types, followed by averaging across observation sites for like classes of each and assignment to the full domain. With the exception of a few sites in the offline calibration results, this approach seemed to work well overall, as evidenced by the independent assessment of 214 locations of T2 and Q2 performance in the coupled run.

![Figure 13](image.png)

**Figure 13.** Mixing diagrams for the 14 July 2006 case study showing the DEF and C06, C07, C08, and C678, along with observations at the ARM-SGP E4 site. Note that DEF and C06 are equal to DEF and SCP from Fig. 4.
(Fig. 7). A next step in this regard is to investigate the
classification at those ARM-SGP sites that saw fluxes
that degraded slightly after calibration to see if the soil
type and land cover representation at those flux towers
was represented accurately by the datasets (STATSGO
and UMD) chosen for this study.

In addressing the final question of when to calibrate,
we found some interesting results that should be taken
into account in future studies. That the calibration in
the wet regime worked nearly as well as the dry regime
parameters suggests that, in order to improve simula-
tions during extremes, the calibration should be able to
improve model bias in a significant fashion (in this case,
during both dry and wet extremes). Clearly, this is not
a one-size-fits-all approach and depends on the season-
ality of a particular location/climate regime, but it also
suggests that the model physics be tested outside of
“average” conditions (even those that include dry and
wet periods within them) in order to maximize LSM
improvement due to calibration. There are many more
experiments that could be performed in terms of period
sensitivity (seasonal, application to average condition
coupled cases, etc.) that will be a part of future research.

Another issue rarely addressed in studies of LSM
calibration is that of the physical realism of the cali-
bbrated parameter values and consideration of what the
values actually represent relative to the default lookup
tables. Santanello et al. (2007) was successful in achieving
both goals of reducing model bias and maintaining pa-
rameter realism among soil hydraulic properties through
the use of pedotransfer functions. Here the parameter
set is so large that it is difficult to ensure or even evaluate

![Fig. 14. Daytime mean evaporative fraction vs PBL height for the simulations in Fig. 13.](image)

<table>
<thead>
<tr>
<th>TABLE 8. Error statistics from Fig. 14 for each of the simulations. Bold values as in Table 3.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Total MAE</td>
</tr>
<tr>
<td>Total RMSE</td>
</tr>
<tr>
<td>Q2 bias</td>
</tr>
<tr>
<td>T2 bias</td>
</tr>
<tr>
<td>N-S efficiency</td>
</tr>
</tbody>
</table>
interparameter consistency and applicability to real-world (or measured) properties, and many LSM parameters are not even observable. More importantly, the primary goal of this study is to evaluate the impact of optimal surface fluxes on coupled forecasts, while realizing that the atmospheric component is not sensitive to actual values of LSM parameters or their consistencies. More rigorous studies focused on generation of “equifinal” solutions (Gupta et al. 1999; Hogue et al. 2005) have shown promise in evaluating physical consistency, but they are more narrowly focused on calibration approaches and internal physics of a particular model.

For most calibration studies in applications outside of hydrology, the ends (i.e., improved flux output) justify the means (i.e., limited parameter realism). However, we can still take a closer look at the evaporative physics in Noah and two of the commonly modified and “tuned” parameters in previous studies. The FXEXP parameter is the exponent for bare soil evaporation in Noah, which is a function of soil moisture and vegetation amount. Lower values of FXEXP increase the bare soil component of Qle for a given soil moisture/vegetation amount, and the default value is 2.0. Table 9 shows the calibrated values from the different period experiments, and there is a definite downward shift in FXEXP due to calibration toward 1.0. In fact, Santanello et al. (2007) modified the FXEXP parameter in their study to be 1.0 because of the semiarid region and inability of Noah to produce enough Qle. The calibration here has acted in the same manner in order to increase Qle to match observations.

The other parameter of interest is part of the evaporative/flux calculations in Noah. CZIL is the Zilitinkevich coefficient relating the roughness length for momentum to the roughness length for heat (Z_{oh}) and the exchange coefficient (C_h). There has been recent work in Noah model development to modify this from its default value of 0.1 to a value dependent on the particular region and/or vegetation coverage (e.g., Mitchell et al. 2004; Chen and Zhang 2009; LeMone et al. 2008, 2010; Trier et al. 2011). Higher values of CZIL decrease Z_{oh}, C_h, and flux magnitudes overall. Table 9 shows the values of CZIL from DEF lookup table of Noah along with calibrated

<table>
<thead>
<tr>
<th>DEF</th>
<th>C06</th>
<th>C07</th>
<th>C08</th>
<th>C678</th>
<th>LeMone et al. (2008)</th>
<th>Trier et al. (2011)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FXEXP</td>
<td>2.0</td>
<td>1.06</td>
<td>1.34</td>
<td>0.969</td>
<td>1.19</td>
<td>—</td>
</tr>
<tr>
<td>CZIL</td>
<td>0.1</td>
<td>0.6</td>
<td>0.6</td>
<td>0.1</td>
<td>0.6</td>
<td>0.5</td>
</tr>
</tbody>
</table>
values from different periods and the prior study estimates. The value has been raised to 0.6 in the calibrations that perform best (C06, C07, C678) versus 0.1 in the DEF and the poor calibration of C08.

These results are consistent with tests of the Noah model over the ARM-SGP domain by LeMone et al. (2008), who found that CZIL should be larger in this region and may also explain why both C06 and C07 perform well for the 2006 case study. Figures 2 and 3 showed that the bulk of improvement in the offline runs were due to reducing Qh in the dry year and reducing Qle in the wet year. Because these are the dominant flux components in each year, the increase in CZIL is at least partially responsible for the improved fluxes and reduction in model bias (consistent results are also found in C678, where both Qle and Qh are improved and reduced). C08 saw no change in CZIL and, therefore, no change in the evaporative physics for the corresponding spinup or coupled run application.

The SPN versus CPL results above also support those of Trier et al. (2008) in terms of consistency in calibrated parameter sets and suggest that the results of Trier et al. (2011) would have shown even greater sensitivity of land–PBL coupling to CZIL if the same modified values were used both in the spinup and coupled runs (their CZIL modifications were applied to the coupled run only). Overall, the calibrated values of both CZIL and FXEXP appear to be physically consistent with previous studies' manual tuning of parameters, and while they by no means guarantee the same for the other 27 parameters involved, they at least suggest some physical consistency and model improvement that produces the right answer for the right reasons.

6. Conclusions

This study examines the impact of LSM spinup and calibration on the land–PBL coupling in regional model forecasts. Sensitivities to dry/wet regimes, period of calibration, and parameter sets were quantified using diagnostics of LA coupling and applied to the NU-WRF coupled modeling system. Key findings from this work include the following.

- Offline calibration using a surface flux network is successful in reducing LSM biases and improving diurnal cycles of Qle and Qh.
- Calibrated parameter sets can improve fluxes and states during both dry and wet regimes and extend their impact to PBL fluxes and ambient weather (T2 and Q2).
- Largest impacts of offline calibration on coupled runs are seen during the dry regime when the turbulent fluxes are larger and atmospheric and precipitation forcing is weak.
- A calibrated spinup by itself can produce more accurate temperature and humidity forecasts, regardless of the parameter sets used in the coupled simulation, though consistency in parameter sets between spinup and coupled runs is critical to improving performance and maintaining physical consistency in both states and fluxes.
- Calibration during primarily dry and/or wet extreme periods corrected more of the inherent LSM bias and led to better coupled predictions in the dry regime.
- Significant variability in hydrometeorological prediction can result from LSM parameter uncertainty but can be reduced using observations and calibration approaches.

These results can be considered preliminary in the sense that there are multiple approaches to offline calibration (objective function, multicriteria, etc.) and the extension of calibrated parameters from a single grid cell to other like sites in that domain. Future work is being planned using LIS-OPT/UE to determine the relative importance of specific parameters in the calibration process and to answer the questions of how many and which parameters are essential to achieving the degree of coupled model improvement seen in this study. Further, these experiments were also designed as a prototype test bed for future satellite missions [e.g., NASA's Soil Moisture Active Passion (SMAP) mission]. Using LIS-OPT/UE, the tradeoffs of data availability versus accuracy and uncertainty in prediction can be quantified systematically. The classification strategy employed in this study (i.e., from single site to fully gridded parameter optimization) also relates to the spatial tradeoffs of satellite sensors, while the period of calibration relates to the satellite overpass return time. In the future, simultaneous development of Earth science technologies (e.g., microwave soil moisture sensors) and methodologies (e.g., thermal evapotranspiration retrievals) will warrant the LIS-OPT/UE approach in assessing the impact of observations on coupled forecasts for both calibration and data assimilation studies alike.

Acknowledgments. This work was supported by NASA's Advanced Information System Technology program (AIST) and the Earth Science Technology Office (ESTO). The NU-WRF team was also instrumental in providing support related to LIS-WRF coupling and a stable and updated version of the coupled system. The data for the MET analysis are from the Research Data Archive (RDA), which is maintained by the
Computational and Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR). The original data are available in dataset number ds337.0 (http://rda.ucar.edu/datasets/ds337.0).

REFERENCES


