A Multimodel Analysis, Validation, and Transferability Study of Global Soil Wetness Products

XIANG GAO AND PAUL A. DIRMEYER
Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

(Manuscript received 1 June 2005, in final form 9 May 2006)

ABSTRACT
Multimodel ensemble forecasting has been shown to offer a systematic improvement in the skill of climate prediction with atmosphere and ocean circulation models. However, little such work has been done for the land surface component, an important lower boundary for weather and climate forecast models. In this study, the authors examine and evaluate several methods of combining individual global soil wetness products from uncoupled land surface model calculations and coupled land–atmosphere model reanalyses to produce an ensemble analysis. Analyses are verified against observations from the Global Soil Moisture Data Bank (GSMDB) with skill measured by correlation coefficient and root-mean-square error (RMSE). A preliminary transferability study is conducted as well for investigating the feasibility of transferring ensemble regression parameters within two specific regions (Illinois and east-central China) and between these two regions of similar climate and land use. The results show that when sufficient validation data are available, one can use a seasonally dependent linear regression to improve the skill of any individual model simulation of soil wetness. Further improvements in skill can be achieved with more sophisticated ensembling methods, such as the regression-adjusted multimodel ensemble mean analysis and regression-adjusted multimodel analysis. However, all the ensembling schemes involving regression usually do not help improve the skill scores as far as the simulation of anomalies of soil wetness is concerned. In the absence of calibration data, the simple arithmetic ensemble mean across multiple soil wetness products generally does as well or better than the best individual model at any location in the representation of both soil wetness and its anomaly. Transferability from one subset of stations from the Illinois or east-central China dataset to another gives satisfactory results. However, results are poor when transferring regression weights between different regions, even with similar climate regimes and land cover. Such an exercise helps us to understand better the virtues and limitations of various ensembling techniques and enables progress toward creating an optimum, model-independent analysis from a practical point of view.

1. Introduction
Global weather and climate models need accurate and reliable initial conditions from which to begin integrations for forecasting or predictability research. Initial conditions usually come from global near–real time analyses (Lambert 1988), global reanalyses (e.g., Kalnay et al. 1996; Gibson et al. 1997), and data assimilation products (e.g., Derber and Rosati 1989; Carton et al. 2000) by using a realistic geophysical fluid dynamics model to ingest an array of observations and propagate the observationally based information across the globe.

Because of deficiencies in observational networks and data assimilation methods, errors in the initial conditions always exist. These errors grow with time as a result of the internal atmospheric variability causing the forecast to bear less resemblance to the true state of the atmosphere. One way to effectively deal with this inherent uncertainty is through ensemble forecasting. This is commonly done by running a single model repeatedly from a set of slightly different, equally viable initial conditions and averaging an ensemble of these forecasts. It has been shown that this technique has a beneficial impact on the skill of the forecasts by reducing noise in model predictions, and an ensemble mean, given a large enough sample, should always outperform individual members in predicting nonlinear systems (Kharin et al. 2001).

With the availability of climate predictions that are
produced by multiple dynamical models, multimodel ensemble forecasting has become increasingly popular. This generally involves using linear methods to produce better predictions by combining several independent forecasts from different numerical models with potentially fundamentally different statistics. Such an approach has been the topic of several studies. Thompson (1977) showed that the mean-square error of forecasts constructed from a linear combination of two independent predictions is less than that of the individual predictions. Fraedrich and Smith (1989) utilized a linear regression method to combine long-range forecasts of the monthly mean tropical Pacific sea surface temperatures from two statistical forecast schemes. Derome et al. (2001) discussed a linear multivariate method for blending climate forecasts produced by two dynamical models. Doblas-Reyes et al. (2000) examined the performance of multimodel climate predictions produced by three general circulation models. They find that the multimodel ensemble offers a systematic improvement for probabilistic forecasts, but improves skill only marginally when verifying the ensemble mean. Krishnamurti et al. (1999, 2000) discuss a “superensemble” approach involving a multimodel linear regression technique for deterministic forecasts and report an apparent systematic improvement in mean-square error for a multimodel forecast over that of the individual model forecasts. In particular, a recent study by Kharin and Zwiers (2002) has drawn much attention. In their study, they consider seven different approaches of combining individual forecasts from a group of climate models to construct an ensemble forecast and apply these methods to an ensemble of 500-hPa geopotential height forecasts derived from the Atmospheric Model Intercomparison Project (AMIP) integrations performed by 10 different modeling groups. They find that the simple ensemble mean performs best over the Tropics, while the regression-improved ensemble mean performs best over the extratropics. The “superensemble” forecast that is obtained by optimally weighting the individual ensemble members does not perform as well as the simple ensemble mean or the regression-improved ensemble mean because the sample size is too small for estimating reliably the relatively large number of optimal weights. All these studies with atmosphere and ocean circulation models have shown that multimodel ensemble forecasting usually produces a forecast more skillful than the individual ensemble members.

Although it is well established that the multimodel ensemble forecasting approach offers a systematic improvement in the skill of climate prediction with atmosphere and ocean circulation models, little such work has been done for the land surface component, an important lower boundary for weather and climate forecast models. One would assume that a way to easily improve upon the performance of individual land surface models is to statistically combine the simulations of multiple models. Also, a similar comparison to that of Kharin and Zwiers (2002) can be made for land surface models in an offline analysis mode to determine whether there is a method of ensembling that is superior to a simple ensemble mean. However, the physics of land surface models is fundamentally different from the fluid dynamics of atmosphere or ocean models, so it is not clear whether such approaches will yield the same results in this context. It is evident that many motivations exist for us to understand clearly the potential utility of multimodel ensemble analysis in the land surface component. First, for the land surface, the choices for observationally based global analyses as the initial conditions are limited, particularly in the case of soil wetness. This is mainly because soil moisture is difficult to measure in situ, few long-term climatologies of any kind exist, and remote sensing techniques are only effective for a thin surface layer that provides little memory to the climate system beyond day one and is thus of little value for prediction (Dirmeyer 1995). On the other hand, there is ample evidence that, over continental areas, weather and climate forecasts can be improved by providing better initialization and description of the subsequent evolution of soil moisture status in forecast models (e.g., Fennessy and Shukla 1999; Dirmeyer 2000; Koster and Suarez 2003; Douville 2004). At present, the most practical method to generate global long-term estimates of soil wetness is to use a land surface scheme (LSS) driven by observed meteorological forcing. Since the result will be heavily influenced by the characteristics of the chosen LSS and no soil wetness product is clearly superior in all situations (Dirmeyer et al. 2004), multimodel ensemble analysis could play a significant role in helping ameliorate the systematic errors of individual LSS and produce a superior estimate to that given by any individual LSS. Such multiproduct ensembling may offer a way to improve global soil wetness analyses with little expense, particularly if the weighting functions found over regions with reliable validation data can be shown to be transferable to other regions. Such improved soil wetness estimates could therefore serve as a baseline observational proxy for use in initializing various LSSs if climatological statistics (mean and variance) of each individual LSS are taken into account properly through a hybridization process as shown in Dirmeyer et al. (2004).

In this study, several methods of combining individual global soil wetness products from uncoupled land surface model calculations and coupled land-
atmosphere model reanalyses to produce an ensemble analysis are examined. The performance of various constructed ensemble analyses is verified against observations from the Global Soil Moisture Data Bank (GSMDB) with skill measured by correlation coefficient and root-mean-square error. Our objectives are to answer such questions as the following: Is multimodel ensemble analysis applicable to the land surface component, as exemplified by soil wetness, with greater skill than any of its individual members? Which global soil wetness products really add value to the ensemble? Which method of combination will lead to maximal skill and produce the most accurate analyses? From a practical point of view, how should we create suitable and consistent soil wetness analyses in reality for initialization of one’s LSS? We also intend to conduct a preliminary examination of the feasibility of transferring ensemble regression parameters within specific areas and between areas of similar climate regime and land use. Such a transferability study should provide us with useful insights on the possibility of achieving improved soil wetness estimates over those areas where observations are not available. The plan of the paper is the following. Section 2 describes the various global soil wetness products examined in this study and in situ observations from the GSMDB. Section 3 presents a general linear regression method that will serve as the basis for constructing several improved regression variants. Regression and multimodel performance is presented in section 4. Section 5 shows the transferability results. Discussion and conclusions are given in section 6.

2. Data description

a. Model estimates

Six model-derived global soil wetness products that span at least 20 yr (1980–99) are examined. All models are built around physically based parameterizations of the surface water balance. Four of the models also include a full representation of the surface energy balance. A synopsis of the characteristics of the included models is given below. For more detailed descriptions of each specific model, please refer to Dirmeyer et al. (2004).

Three global soil wetness products in this study come from uncoupled land surface model calculations. The simplest model of the set is that of Willmott and Matsuura (2001), referred to hereafter as W&M. It is based on a single-layer bucket-type soil hydrology with a 150-mm capacity. The monthly mean precipitation used to drive the model is produced with Global Historical Climatology Network (GHCN, version 2) and Legates and Willmott’s (1990b) station records. The number of GHCN stations used and stations taken from the Legates and Willmott archive are 20,599 and 26,858, respectively. Station averages of monthly precipitation are interpolated to a 0.5° × 0.5° grid with the interpolation algorithm performed based on the spherical version of Shepard’s distance-weighting method and with the digital elevation model (DEM)-assisted Legates and Willmott archive (C. J. Willmott et al. 1998, personal communication) employed as the background field for the climatologically aided interpolation (CAI; Willmott and Robeson 1995). The dataset is monthly, spans 50 yr (1950–99), and is at a spatial resolution of 0.5°.

A similar global product has recently been developed at the Climate Prediction Center (CPC; Fan and Van den Dool 2004) based on the algorithm of Huang et al. (1996). The monthly precipitation input to CPC is derived by interpolation of gauge observations from over 17,000 stations collected in the GHCN version 2, and the Climate Anomaly Monitoring System (CAMS) datasets (Chen et al. 2002). They choose optimum interpolation (Gandin 1965) algorithm and perform the interpolation for anomaly separately from interpolation of climatology. The model is also based on a single-layer bucket hydrology, but with a capacity of 760 mm. The CPC data cover the period from 1948 to the present. The data are updated daily in near–real time and are available at 0.5° resolution. Daily data have been averaged to monthly for this study.

A modified version of the simplified Simple Biosphere LSS (SSiB; Xue et al. 1991, 1996; Dirmeyer and Zeng 1999) was integrated in an offline mode from 1976 to 1999 to produce the Global Offline Land surface Dataset (GOLD; Dirmeyer and Tan 2001). The meteorological forcing used to drive the SSiB model comes from the National Centers for Environmental Prediction (NCEP) reanalysis of Kalnay et al. (1996), blended with observations where available. GOLD uses a number of regional precipitation datasets to provide a regional correction to the monthly mean rainfall of the reanalysis. These include the rainfall data of Higgins et al. (1996) at 2.5° spatial resolution over the United States for the entire time period, the gridded data of Webber and Willmott (1998) at 1° spatial resolution over South America until 1990, and the station data of Singh et al. (1992) over India for the period of 1971–90. The CPC Merged Analysis of Precipitation (CMAP; Xie and Arkin 1997) dataset at 2.5° spatial resolution, spanning the period of 1979–99, is also used to provide a correction to the monthly mean rainfall of the reanalysis over these three regions beyond the periods of those data mentioned above, and over areas without
regional precipitation datasets. There are three soil layers in the GOLD model: a surface layer of 50-mm depth, a root zone, and a deep recharge zone. The depths of the two lower layers vary spatially, but the total soil column is nowhere less than 1 m. Monthly data at 1.875° resolution are used for this study.

The other three model-based soil wetness products were all produced by coupling to a global atmospheric circulation model (GCM) integrated in data assimilation mode from operational atmospheric reanalysis efforts. The European Centre for Medium-Range Weather Forecasts (ECMWF) produces a 40-yr global reanalysis (ERA-40; Simmons and Gibson 2000), spanning the period from September 1957 to the present. The reanalysis is at a spatial resolution of 1.125° with monthly volumetric soil wetness data for four soil levels provided. There are two reanalysis products from NCEP. The first is the NCEP–National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996). The resolution of the reanalysis is T62, or 1.915° latitude × 1.875° longitude, and monthly data are available from the late 1940s to present. Another reanalysis estimate is from the NCEP–Department of Energy (DOE) reanalysis for the second Atmospheric Model Intercomparison Project (AMIP-II) period (Kanamitsu et al. 2002). The NCEP–DOE reanalysis spans the period 1979–present. The resolution of this reanalysis is same as that of NCEP–NCAR. Daily data are averaged to monthly means for this dataset. Both NCEP reanalysis products share the same two-layer vertical structure of the soil with a surface layer of 100-mm depth and a lower layer that reaches 2 m below the surface. The precipitation in all three reanalysis products is purely the result of the reanalysis model, but the NCEP–DOE reanalysis incorporates observed precipitation (Xie and Arkin 1997; Xie et al. 2003) rather than model precipitation to drive soil moisture in the LSS, and corrects a number of errors found in the original NCEP–NCAR product.

We have paid particular attention to describing the precipitation data used in each of these products because of the paramount role played by precipitation in determining soil wetness. All of the gauge-based precipitation analyses suffer from common problems, such as poor or uneven sampling over many parts of the globe, temporal inhomogeneity in historical records, and uncertainty in gauge undercatch due to wind and evaporation effects, especially where snowfall is prevalent. Attempts to deal with these shortcomings can lead to differences among gridded observational datasets in poorly sampled regions, such as those documented for CMAP over western equatorial Africa by Yin et al. (2004). The reanalyses report model precipitation during an interval of the global model integration some hours beyond the analysis time. These are actually short-term forecasts of precipitation that do not contain any information from precipitation observations, are not well constrained by other atmospheric observations because of the intervening role of the model parameterizations in generating the precipitation forecasts, and tend to be heavily influenced by the “spinup” of the atmospheric model from a rain-free initial condition following each analysis interval. As a result, reanalysis precipitation is prone to biases and is considered to be an undependable quantity in reanalyses (Kalnay et al. 1996). However, the pattern of interannual variations in precipitation may yet be well represented, provided the reanalysis model can characterize its relationship to observed and assimilated variations in the general circulation and atmospheric thermodynamics. Soil wetness in the NCEP–DOE reanalysis is controlled by observed precipitation (Lu et al. 2005); thus, it may behave more like the uncoupled LSS products. This comparison should reveal whether this is the case.

### b. Observations

The GSMD contains the most complete collection of in situ soil moisture observations in terms of spatial and temporal coverage (Robock et al. 2000). This collection of soil moisture measurements, mostly made gravimetrically and over grass vegetation, covers a large variety of climatic regions, including the United States, China, India, Mongolia, and the former Soviet Union (FSU). Soil moisture observations for over 600 stations with records from 11 to more than 20 yr are assembled, quality controlled, and made public to the whole science community. These datasets have been widely used to evaluate simulations of soil moisture from climate models, to study the spatial and temporal scales of soil moisture variations, to validate satellite-retrieved soil moisture, and to design new soil moisture observation networks. In this study, we focus on the station data from Illinois and from east-central China. Stations in east-central China are chosen to conduct a transferability study of ensemble regression parameters with those in Illinois because of the presumed similar climate regime (comparable mean temperature and precipitation) and vegetation (cereal crops) in the two regions.

The data over Illinois, which span January 1981 to June 2004, are the most complete and well documented (Hollinger and Isard 1994). The measurements are made at 19 stations for 12 levels (10, 30, 50, 70, 90, 110, 130, 150, 170, 190, and 200 cm). Field capacity data are available for all the layers as well. The vegetation at all stations is grass, except for one station with bare soil
measurements (Dixon Springs—bare) and at the same location as a grass-covered station (Dixon Springs—grass). Among all the stations with grass, one station (Topeka) has much different soil properties (loamy sand) in comparison with others (silt loam) (Hollinger and Isard 1994), and another station (Fairfield) is a rather new site with relatively short time series (B. Scott 2006, personal communication). These two stations together with bare soil station are withheld from our analyses, leaving 16 stations for this study. The China dataset consists of 43 stations for the period 1981–91. Gravimetric soil moisture observations are taken three times (8th, 18th, and 28th) per month at 11 levels (5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 cm). The vegetation present in the network is categorized as either agricultural or grassland. We have chosen eight stations located within east-central China for this study (Fig. 1).

**c. Preprocessing of the datasets**

Some preprocessing has been performed in order to standardize various global soil wetness products and in situ observations. Each dataset is averaged if necessary to get monthly mean values of soil moisture. The monthly mean is then converted to a 0–1 scale soil wetness index (SWI), where 1 represents saturation, and 0 represents complete desiccation. If the monthly mean value is expressed in units of mass or depth in each layer, these values are normalized by the local saturated capacity of each layer to obtain a SWI. In the case of volumetric soil wetness, scaling by porosity is performed to give a SWI.

Where a product has multiple vertical soil layers, an approximation of a root zone (column available) SWI extending at least 1 m down is calculated and included in the following analysis. The surface SWI represented by the uppermost layer is also analyzed, but the results are not shown here. For the NCEP reanalysis products, the root zone includes both layers (down to 2 m). For the ERA-40, the top three layers (down to 1 m) are used. Layer depths below the surface vary in the GOLD product, but the top two layers are always used, which descend to a range of 0.5 ~2.5 m. The variations in soil column depth among models will lead to differences in the timing of subsurface soil moisture anomalies (lag increasing with depth). However, it is assumed that averaging to monthly means eliminates most of this potential inconsistency.

In situ observations from the GSMDB generally have a much higher vertical resolution than the model products used in this study. Similarly, the soil moisture in the column down to 1 m is used to estimate the column-available SWI. For the data over Illinois, the daily measurements for the top five layers plus half of the amount from the sixth layer are combined and averaged whenever the data are available to get monthly mean values for the column-available soil moisture. Available field capacity data are then used to produce the column-
available SWI. For the data over China, the measurements for all the layers are combined to give the column-available soil moisture. Because of the lack of the soil characteristics information at the China stations, the maximum instantaneous values of soil moisture at each layer over the 11-yr period are assumed as the estimated field capacity and used to derive a SWI for that layer.

We try to choose the period of the study for each region with maximal overlap between the model products and observational dataset. Both W&M and GOLD data cover the period until 1999, rendering our analyses to stop in 1999. Considering that the Illinois data from the first three years (1981–1983) are quite different from those post-1983 as a result of instrumentation changes (http://climate.envsci.rutgers.edu/soil_moisture/illinois.html), only 16 yr of data (1984–1999) have been used. In China, there are 11 yr of data available (1981–1991). For a specific station, if more than 75% of the observational data are missing during the period of record for that dataset, that station is excluded from the validation exercise.

3. Method

Kharin and Zwiers (2002) described a general linear regression method for an ensemble of forecasts produced by $M$ models that will serve as the basis for constructing several improved forecasts. An arbitrary linear combination of these forecasts is given by

$$F(t) = a_0 + \sum_{i=1}^{M} a_i X_{i}(t).$$

(1)

where $X_{i}(t), i = 1, \ldots, M$ denote an ensemble of forecasts produced by $M$ models at a fixed location, and $F$ represents the multimodel regression forecast. The $M + 1$ coefficients $a_i, i = 0, \ldots, M$ are subject to some constraints. Within the bounds of those constraints, coefficients may be chosen to minimize the mean-square error of $F$ against the observations. Depending on the constraints and the extent to which the coefficients in Eq. (1) are adjustable, several variants of this general regression approach could be derived as shown in Kharin and Zwiers (2002) and will be examined in this study as well. We consider the following regression variants:

- The arithmetic ensemble mean forecast $A$ is the simplest approach to the multimodel ensembling. This variant is obtained by setting the coefficients $a_0$ to 0 and $a_1, \ldots, a_M$ to $1/M$ [Eq. (2)]. No coefficients need to be estimated:

$$A(t) = \frac{1}{M} \sum_{i=1}^{M} X_{i}(t).$$

(2)

- The regression-improved individual forecast $R_i$ (Kharin and Zwiers 2002) is obtained by linearly regressing the $i$th simulation against the observations. This is equivalent to setting all coefficients to 0, except for $a_i$ and $a_0$ [Eq. (3)]:

$$R_i(t) = a_0 + a_i X_{i}(t).$$

(3)

- The regression-improved multimodel ensemble mean forecast $R_{EM}$ (Kharin and Zwiers 2002) is obtained by linearly regressing the multimodel ensemble mean against the observations. This is equivalent to constraining the coefficients $a_i = a, i = 1, \ldots, M$, to be equal [Eq. (4)]:

$$R_{EM}(t) = a_0 + a \sum_{i=1}^{M} X_{i}(t).$$

(4)

- The regression-improved multimodel forecast $R_{all}$ (Kharin and Zwiers 2002) is obtained by fitting Eq. (1) to the observations with no constraints on the regression coefficients.

For atmosphere and ocean models, skill improvement of these regression variants is achieved by reduction of noise, thus enhancing the ratio of signal to noise in the simulations. For land models, noise is not a major issue, since the models behave in a more deterministic fashion than fluid dynamical systems. The gain in skill is realized by the correction of systematic errors and pathological behaviors in individual LSSs, which arise from poor calibration or an incomplete representation of physical processes.

In this study, we apply those regression variants described above for soil wetness time series and their anomalies. As we are not making any forecast but only apply regression and we are not certain yet if regression will definitely yield better skill, we use “analysis” instead of “forecast” and “adjusted” instead of “improved” hereinafter as the terms to describe each regression variant, such as arithmetic ensemble mean analysis, regression-adjusted individual analysis, regression-adjusted multimodel ensemble mean analysis, and regression-adjusted multimodel analysis. The performance of the regression variants is evaluated against observations from the GSMDB using correlation coefficient and root-mean-square error (RMSE) metrics. We take into account the possible variations within the annual cycle by allowing the regression coefficients to depend on the calendar month. The skill scores are estimated with a cross-validation procedure in which one calendar month of a year is repeatedly withheld from all the years of data for that calendar month and the regression coefficients are estimated from the retained years (von Storch and Zwiers 1999). The cross-
validation procedure is implemented for all the regression variants. For comparison purposes, $R_{all}$ is also performed without a cross-validation procedure. This is because, given the relatively short span of observational data (as well as the consequences of missing data) and an ensemble of six models in $R_{all}$, there are few degrees of freedom in terms of the total number of fitted coefficients in the multimodel regression technique.

4. Regression and multimodel performance

In this portion of the study, we have utilized only Illinois soil moisture data to evaluate the performance of various regression variants.

a. Total field

Figure 2 shows skill scores of the original soil wetness model products against the regression-adjusted individual analysis $R_i$ based on the observations of 16 Illinois stations from the GSMDB. Also shown are skill scores from station/grid-averaged dataset. As we can see, linearly regressing each model improves skill scores substantially. The gain in correlation is quite evident. However, the performance varies substantially from one model to another and from one station to another. The NCEP–DOE reanalysis, ERA-40, and CPC show the greatest gain in correlation (0.0 ~ 0.3). The unregressed NCEP–DOE product previously has been shown to be poor in this metric (Dirmeyer et al. 2004). Lu et al. (2005) speculate that this may be the result of the highly empirical nature of the soil moisture adjustment employed. GOLD and NCEP–NCAR re-analysis have moderate gain in correlation (0.0 ~ 0.1), while most stations for W&M indicate some loss or no gain in correlation (−0.15 ~ 0.0). The reduction of RMSE can be clearly observed with RMSE generally below 0.15. GOLD and NCEP–NCAR reanalysis show less reduction in RMSE (0.0 ~ 0.1), while other model products have a greater reduction (0.0 ~ 0.3). All the characteristics present at the individual stations can also be observed for the station-averaged dataset. However, station-averaged data clearly outperform each individual station in terms of correlation and RMSE. This is understandable as station-averaged data have less noise and random error. These results indicate that simple linear regression does improve the skill of any individual model simulation of soil wetness time series. The only requirement is the availability of the observations to perform such regression.

For multiple simulations with the same atmospheric general circulation model (AGCM) in weather and climate prediction, it can be shown that for a reasonably large ensemble of simulations with similar statistics, the mean-square error of an arithmetic ensemble mean simulation is smaller than that of a typical ensemble member. In this study we also attempt to examine the arithmetic ensemble mean analysis ($A$) but with an ensemble of different LSSs. Figure 3 shows the comparison of skill scores of the original soil wetness model products and the arithmetic ensemble mean for 16 Illinois stations and the station/grid-averaged dataset. As can be seen, the arithmetic ensemble mean across models without any processing gives skill comparable to that of the best individual model, with scores of 0.7 to 0.8 for correlation and 0.1 to 0.2 for RMSE where soil
wetness ranges from 0.0 to 1.0. One or more individual models may still outperform the arithmetic ensemble mean for a specific station. W&M demonstrates apparent superiority of correlation for most stations, while GOLD and NCEP–NCAR give consistently better RMSE in comparison to the ensemble mean. Therefore, there is no guarantee of a consistently better simulation from the ensemble mean compared to any individual model. Such a characteristic arises from the fact that we are averaging together the simulations of different models with fundamentally different statistics. When the same model is used to produce each ensemble member, there is no reason to expect one realization to consistently outperform the ensemble mean. Despite this fact, no individual model here exhibits clear superiority of skill scores in all situations. Thus, a simple arithmetic ensemble mean across multiple models could still serve as a potentially useful analysis to ameliorate the systematic errors of individual models if there is not enough local observational data to justify a correction by statistical methods.

Comparisons of skill scores for the arithmetic ensemble mean analysis \((A)\) and more complex approaches \((R_{\text{EM}}\) and \(R_{\text{all}}\)) are shown in Fig. 4. For verification purposes, skill scores for \(R_{\text{all}}\) calculated without cross-validation are included as well. Regression on the multimodel ensemble mean \((R_{\text{EM}})\) slightly outperforms the arithmetic ensemble mean with comparable correlation for most of the stations but with a large reduction in RMSE. On the average, RMSE decreases about 0.1. Some further improvement in skill scores can be achieved when a regression is performed without cross-validation for all the adjustable coefficients of individual models \((R_{\text{all}})\). Such improvement over \(R_{\text{EM}}\) is reflected by an average increase of 0.12 for correlation and an average decrease of 0.02 for RMSE, respectively. However, this is not the case when regression of \(R_{\text{all}}\) is performed with cross-validation. At indi-
Individual stations, skill scores for $R_{\text{all}}$ are generally lower than those for $R_{\text{EM}}$. In comparison with the arithmetic ensemble mean, $R_{\text{all}}$ has much worse correlation but better RMSE. It is expected that $R_{\text{all}}$ without cross-validation has much higher skill scores than with cross-validation as the skill is assessed with the same data that are used to train the regression model (i.e., verification data are not independent of the training data). The relatively lower skill scores for $R_{\text{all}}$ with cross-validation than those for $R_{\text{EM}}$ may arise from the fact that the number of regression coefficients to be estimated is large relative to the sample size, resulting in overfitting that causes overly optimistic estimates of skill. As documented in Kharin and Zwiers (2002) as well, the fitted model in such circumstances performs poorly on independent data because it has adapted itself to the available data in the training period. We notice that for some stations there are excessive missing observations, which makes the regression untenable if not actually unsolvable for $R_{\text{all}}$ with cross-validation in contrast to $R_{\text{all}}$ without cross-validation.

In summary, among various multimodel regression variants examined here ($A$, $R_{\text{EM}}$, $R_{\text{all}}$), the best skill may be had when each model has its own optimal calibrated weight, assuming that very large training datasets are available to estimate reliably all the optimal weights required for the $R_{\text{all}}$ approach. However, a sufficiently long and complete observational dataset is needed to perform such regression and achieve a consistent degree of improvement in skill scores. This condition exists in very few locations around the world. In the absence of suitable regressions, the arithmetic ensemble mean analysis ($A$) across multiple models generally does as well as or better than the best individual model at any location.
b. Interannual variability

The key to usefulness of any global soil wetness product for improving the initialization of weather and climate forecasts lies in its ability to represent anomalies (mean annual cycle removed). Simulation of anomalies by $R_i$, $R_{EM}$, and $R_{all}$ in this study is implemented through regression on the anomalies of SWI. Figure 5 shows skill scores of the anomalies of the original soil wetness model products against those of the regression-adjusted individual analysis $R_i$. In contrast to the results for the total field, linear regression on the anomaly of each model generally does not help improve skill scores, with lower correlation and slightly higher RMSE. When cross-validation is not applied, the regression-adjusted analysis is usually better than the original correlation, and RMSE is slightly lower (not shown). A further investigation on regressions of each calendar month suggests the sample size may be too small to allow a reasonable cross-validated regression on the weaker and less systematic signal of the anomalies. A single outlier or displaced data point can dramatically change the resulting regression. This is generally not true for regression on the total field, which for most locations is dominated by the annual cycle of soil wetness, and in which interannual variations are of secondary magnitude. Barnston and Van den Dool (1993) demonstrate that correlation skill score degeneracy may occur in regression-based cross-validation. Such degeneracy may result in degradation of skill scores in this situation. There is evidence that the distribution of the cross-validation regression-adjusted correlation against the “full sample” correlation in Fig. 5 resembles the case of degeneracy with a sample size of 16, plus some scattered skewing due to the effects of outlier (like Figs. 1a and 2 in Barnston and Van den Dool 1993). However, it is difficult to discern how the specific elements discussed in Barnston and Van den Dool may apply in our more complicated approach as our cross-validation regression-adjusted correlation has taken into account the possible variations within the annual cycle. In addition, the effect of degeneracy could become more severe when the requirement for statistical significance for the sample size is not met. Therefore, we think the failure of regression-adjustment in anomaly is still primarily an issue of sample size—with a sufficiently large sample (many decades) and the significantly large correlations we see in our results, the effects of CV degeneracy would be minimized.

Comparisons of skill scores for the anomalies of the original soil wetness model products and their arithmetic ensemble mean are shown in Fig. 6. Again, the arithmetic ensemble mean across the anomalies of multiple models generally gives skill comparable to that of the anomalies of the best individual model. For correlation, the ensemble mean ranks first or second mostly and ranks as low as third only twice. Among the individual models, W&M, ERA-40, and GOLD exhibit clear superiority of skill and can all serve as a candidate of the best individual model. All these features are consistent with what we have observed for the total field. This further confirms that the degradation of skill scores for the anomalies of individual models by regression (Fig. 5) can be attributed to small sample
size when performing a representative cross-validated regression.

We also calculate the skill scores for the anomalies of more sophisticated regression schemes (\(R_{EM}\) and \(R_{all}\)). Comparison of correlation for the anomalies of \(R_{EM}\) and \(R_{all}\) with that for the arithmetic ensemble mean is shown in Fig. 7. (RMSE is not shown as the anomalies of \(R_{EM}\) and \(R_{all}\) have exactly the same RMSE as the total field.) As can be seen, the arithmetic ensemble mean has the highest correlation in comparison with \(R_{EM}\) and \(R_{all}\) with cross-validation. This is different from what we observe with the total field. It is evident that regression on the anomalies does not help improve the correlations, even for the scheme of \(R_{EM}\). Overall, correlations for \(R_{all}\) with cross-validation are lower than that for \(R_{EM}\). As we mentioned above, this could result from the small sample size relative to the number of regression coefficients to be estimated. Nevertheless, we can conclude with some confidence that for all locations but those with ample in situ data, the best scheme will be the simple arithmetic ensemble mean across multiple models.

c. Relative contributions of the models to \(R_{all}\)

The exercise of the regression-adjusted multimodel analysis (\(R_{all}\)) with global soil wetness products is a useful one to assess quantitatively if there exists an individual model product that consistently performs best and is best suited to using as a proxy for global observations for the purpose of initializing weather or climate models. With the increasing number of multimodel experiments and intercomparisons, there needs to be a consistent way to initialize the land surface state across many models. However, soil moisture is usually not directly transferable among weather and climate models because different models have different operating ranges and different sensitivities of evaporation and runoff (Koster and Milly 1997). If such a “consistently superior” model does exist, we can treat it as the baseline product and translate soil wetness to any target.
model using standard normal deviates to maintain the statistical characteristics (the mean annual cycle and the variance) of the target model (Dirmeyer et al. 2004).

Here we define the relative contribution of an individual model in $R_{\text{all}}$ as the following:

$$W_k = |a_k^* \sigma_k|,$$

(5)

where $W_k$ and $a_k$ are the relative contribution and the regression coefficient of model $k$ in $R_{\text{all}}$, respectively, for a specific month and station, and $\sigma_k$ represents the interannual standard deviation of model $k$ calculated from all available data for the given model product. Note that any given model may contribute strongly with either a positive or negative value of $a_k^*$; anticorrelations also contribute to the regression $R_{\text{all}}$. In the following sections (including transferability study), our analyses with regard to $R_{\text{all}}$ will be based on the calculations without cross-validation only.

Figure 8a shows the relative contribution averaged over all the Illinois stations for each month. As can be seen, there is much variation in each model’s relative contribution among the months. There is no consistently “good” or “bad” model. Each model dominates some months and trails in others. GOLD contributes most for the first half of the year, but W&M and GOLD contribute equally strongly for the second half of the year. The results from the station/grid-averaged dataset also confirm what we observe above, except that ERA-40 shows consistently large contributions during the second half of the year as well (Fig. 8b). Note that all the models appear to have large relative contribution for November in Fig. 8a, but this is not obvious for the station/grid-averaged dataset in Fig. 8b. Further analyses reveal that the relative contribution averaged over all the Illinois stations could be strongly influenced by a small number of specific stations. Another contributing factor could be that November has the fewest valid station reports of any month. However, this does not hold for the station/grid-averaged dataset. The relative contribution averaged over all the months for each station is shown in Fig. 8c. As can be seen, there is much variation in each model’s relative contribution among the stations as well. For some stations, each model contributes similarly, while for other stations there are marked differences. It appears that GOLD and W&M contribute robustly to most stations.

5. Transferability

We have shown that the performance of any individual model or multimodel simulation of the soil wetness time series can be greatly improved if the regression coefficients of various regression variants are appropriately estimated on the basis of calibrations with observations. In data-sparse areas, however, it is highly unlikely that there are enough observations for calibration. Therefore it is very important to understand the feasibility of transferability for parameters of various regression variants. In other words, can we calibrate regression parameters over the data-rich areas and apply the calibrations for some other data-sparse areas in the world so that similar performance of soil wetness simulation may be expected? Wood et al. (1998) pointed out that the problem of how to transfer information from calibrated areas to other areas remains unresolved. Huang et al. (2003) illustrated a framework for successfully transferring the b parameter of the
Fig. 8. Relative contribution to $R_{\text{swi}}$ of individual soil wetness products (a) averaged over all the Illinois stations for each month; (b) for area (station/grid) averaged SWI; and (c) averaged over all the months for each Illinois station for column-available SWI.
Variable Infiltration Capacity (VIC) land surface model from data-rich areas to data-sparse areas. In this study, we choose eight stations in Illinois and in east-central China from the GSMDB to examine the feasibility of transferring ensemble regression parameters within each specific region and between them (Fig. 1). Eight stations are divided into two groups for transferability study within each specific region (Fig. 1), while all of them are used for transferability study between the two regions. The stations in each group are selected so as to be spatially distributed over the region and with similar stations going into different groups. These two regions are chosen because they have similar climates (mean temperature and precipitation) and vegetation (cereal crops) and also have long-term observational soil moisture data for verification.

Figure 9 shows the skill scores of data averaged from four Illinois stations (group 1 in Fig. 1) with locally calibrated regression parameters (x axis) against those of the same dataset with regression parameters calibrated using data averaged from another four Illinois stations (group 2; y axis). As can be seen, we are able to transfer the calibrations quite well within the Illinois region for almost all the ensemble schemes with correlation and RMSE fairly well preserved. However, there appears a slight degradation in skill for the $R_{\text{all}}$ case. Transfer of regression parameters from group 1 to group 2 presents similar results (not shown). In summary, within the Illinois region, we seem to be able to mutually transfer regression parameters quite well between two groups of stations.

Results from the transfer of regression parameters between two groups of four east-central China stations are shown in Fig. 10. As in the Illinois case, transferring the calibrations works reasonably well from one group of four stations to another for almost all the schemes, except that there appears to be a relatively large loss in skill for the $R_{\text{all}}$ case. However, all the skill scores in east-central China appear to be lower than those in Illinois. Note that correlation and RMSE of the original soil wetness model products are also lower for stations in east-central China than for those in Illinois. We have also investigated transfer of regression parameters within both regions (Illinois and east-central China) with different sample sizes and the selection of different stations. The resulting skill scores are found to be slightly dependent on both factors, but the general conclusions remain the same. Selection of the particular stations for coefficient transfer appears to be especially important when the sample size of the dataset is small and variability among stations is relatively large.

Figure 11 shows skill scores for the transfer of regression parameters derived from eight stations in Illinois and in east-central China when calibrated at stations in the same region (x axis) or at eight stations in the other region (y axis). The results are not satisfactory. When the calibrations from the China dataset are applied to the stations in Illinois, there appears to be quite a large loss in correlation (0.1 ~ 0.6 decrease). RMSE is also strongly degraded by the transfer with the resulting score values of 0.1 ~ 1.4. The performance appears to be slightly worse for correlation, but slightly better for RMSE when we transfer the calibrations from the stations in Illinois to the east-central China region. Among
all the schemes, $R_{\text{all}}$ seems to degrade most. There is a suggestion in the top-left panel of Fig. 11 that the correlations may be worse when transferring products based on observed precipitation than when a model precipitation is driving soil moisture. One might expect transferability to be poorer where observed precipitation is used because observed precipitation is inherently noisier in a statistical sense than grid-box-averaged model precipitation. But this phenomenon is not evident when transferring from Illinois to China (lower-left panel of Fig. 11), only from China to Illinois. A larger set of transfer exercises would be necessary to determine if there is a systematic difference in behavior between the products driven by observed precipitation and the reanalyses. In summary, there exists some transferability in this case, but the results are not comparable to those within the specific region. It appears that caution has to be taken when transferring ensemble regression parameters between different regions, even with presumed similar climate regimes and land use. This also verifies that in the derivation of a multimodel ensemble analysis on a large scale, we will not be able to safely attempt any technique more sophisticated than a simple average in the absence of observational data to obtain a statistically robust regression.

In this study, all the analyses are also performed for the surface SWI. Our results indicate that the surface SWI has quite similar characteristics for the performances of multimodel regression variants and transferability of regression parameters to those of the column-available SWI. Thus, all of the conclusions found in the
column-available SWI are also applicable to the surface SWI.

6. Discussion and conclusions

In this paper, we have examined several methods of combining global soil wetness products from uncoupled land surface model calculations and coupled land-atmosphere model reanalyses in order to produce an ensemble analysis. The performance of various versions of monthly simulations of soil wetness, as measured by time series correlation coefficients and RMSE, have been evaluated based on in situ measurements. A preliminary transferability study has also been conducted to examine the feasibility of transferring tuned parameters between two specific regions (Illinois and east-central China) of similar climate regime and land use. We have two goals in this research. One is to determine whether there is a method of statistically correcting or combining soil wetness products to produce a consensus product whose performance is superior to any individual product. The second is to determine whether weighting coefficients calibrated over regions with reliable validation data could be transferable to other regions lacking observations. Such an assessment may offer a way to improve global soil wetness analyses at little expense.

Our results indicate that where validation data are available, a seasonally dependent linear regression can be used to improve substantially the skill scores of any
individual model simulation of soil wetness time series. The arithmetic ensemble mean across multiple models generally gives a skill comparable to or better than that of the best individual model, although one or more individual models may still outperform the mean for a specific location. Further improvements in skill can be achieved with more sophisticated ensembling methods. A regression on the multimodel mean \( R_{\text{EM}} \) performs better than the arithmetic ensemble mean, while a comparable correlation but much lower RMSE can be had when each model has its own optimal calibrated weight \( R_{\text{all}} \). However, the performance of \( R_{\text{all}} \) is generally not as good as that of \( R_{\text{EM}} \) in our study. This is consistent with the results of Kharin and Zwiers (2002). It seems that the sample size evidently is too small in current observational soil moisture datasets to estimate reliably the relatively large number of optimal weights required for the \( R_{\text{all}} \) approach.

In terms of simulation of the anomalies, the arithmetic ensemble mean has skill comparable to that of the anomaly of the best individual model. However, none of the linear regression methods improves skill scores, but instead makes them worse. The main reason for the poor performance in simulation of the anomalies by regression is that the sample size is too small to perform a reasonable cross-validated regression on the weaker and less systematic signal left with the anomaly fields. This differs from our results with the total field, which is dominated by the annual cycle of soil wetness for most locations, and in which interannual variations are of secondary magnitude.

We have also defined a means of assessing quantitatively the relative contribution of each global soil wetness product in the regression-adjusted multimodel analysis \( R_{\text{EM}} \). This exercise is useful to help determine if there exists an individual soil wetness product that consistently performs better and would be best suited to using as a proxy for global observations for the purpose of initializing weather or climate models. Our results indicate that there is in fact much variation in the relative contribution of the individual global soil wetness products across months and stations. There is no consistently “best” or “worst” product. Each product dominates some months or contributes to some stations. This result gives further justification to the multimodel analysis approach.

Transferability of regression weights from one subset of Illinois or east-central China station data to the other generally gives satisfactory results with skill scores of the regression-adjusted individual analysis \( R_i \) and the regression-adjusted multimodel ensemble mean analysis \( R_{\text{EM}} \) nearly unaffected and those of the regression-adjusted multimodel analysis \( R_{\text{all}} \) showing a small loss in Illinois and a relatively large loss in east-central China. However, when transfer of regression weights is performed between regions (e.g., from Illinois to east-central China), the resulting skill is much worse than for site-specific calibrations. This confirms that we will not be able to apply any technique more sophisticated than a simple arithmetic ensemble mean in producing the best multimodel analysis on a large scale.

Although better skill can be obtained with more sophisticated ensemble schemes involving regression, one of the biggest difficulties with those schemes lies in the general requirement of the availability of large observational datasets to perform a statistically stable regression. This is feasible only in data-rich areas, which for soil wetness are few. The lack of transferability for ensemble regression parameters between different regions further prevents the implementation of those schemes on a large scale. On the other hand, the arithmetic ensemble mean generally does as well or better than the best individual model in most of the situations. Since the best model varies from place to place and time to time, the multimodel mean is the best overall estimate for soil wetness. The simplicity of this approach and to the fact that we do not need to estimate any coefficients makes it easily applied everywhere on the globe. Therefore, the multimodel arithmetic ensemble mean is our best current hope to achieve improved global soil wetness analyses. Methods to infer local soil properties from the observed temporal behavior of soil wetness (especially from remote sensing) may be able to extend improvements in the simulation of soil wetness beyond the point where practical transferability fails (e.g., Liou et al. 1999; Njoku and Li 1999; Parada and Liang 2004). Despite the poor performance of the transferability exercise, this study helps us better understand the virtues and limitations of various multimodel ensembling techniques. It also enables progress toward creating an optimum, model-independent analysis from a practical point of view.

Acknowledgments. The authors thank the providers of the various soil wetness products, and T. DelSole for discussions on the statistical ramifications of our results. This work is supported by National Aeronautics and Space Administration Grant NAG5-11579.

REFERENCES


