Evaluation of the Reanalysis Products from GSFC, NCEP, and ECMWF
Using Flux Tower Observations

MARK DECKER
Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona, and Climate Change Research Centre,
University of New South Wales, Sydney, New South Wales, Australia

MICHAEL A. BRUNKE, ZHUO WANG, KOICHI SAKAGUCHI, AND XUBIN ZENG
Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona

MICHAEL G. BOSILOVICH
Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 29 December 2010, in final form 1 October 2011)

ABSTRACT

Reanalysis products produced at the various centers around the globe are utilized for many different scientific
endeavors, including forcing land surface models and creating surface flux estimates. Here, flux tower obser-
vations of temperature, wind speed, precipitation, downward shortwave radiation, net surface radiation, and
 latent and sensible heat fluxes are used to evaluate the performance of various reanalysis products [NCEP–
NCAR reanalysis and Climate Forecast System Reanalysis (CFSR) from NCEP; 40-yr European Centre for
Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) and ECMWF Interim Re-Analysis
(ERA-Interim) from ECMWF; and Modern-Era Retrospective Analysis for Research and Applications
(MERRA) and Global Land Data Assimilation System (GLDAS) from the Goddard Space Flight Center
(GSFC)]. To combine the biases and standard deviation of errors from the separate stations, a ranking system is
utilized. It is found that ERA-Interim has the lowest overall bias in 6-hourly air temperature, followed closely by
MERRA and GLDAS. The variability in 6-hourly air temperature is again most accurate in ERA-Interim.
ERA-40 is found to have the lowest overall bias in latent heat flux, followed closely by CFSR, while ERA-40
also has the lowest 6-hourly sensible heat bias. MERRA has the second lowest and is close to ERA-40. The
variability in 6-hourly precipitation is best captured by GLDAS and ERA-Interim, and ERA-40 has the lowest
precipitation bias. It is also found that at monthly time scales, the bias term in the reanalysis products are the
dominant cause of the mean square errors, while at 6-hourly and daily time scales the dominant contributor to
the mean square errors is the correlation term. Also, it is found that the hourly CFSR data have discontinuities
present due to the assimilation cycle, while the hourly MERRA data do not contain these jumps.

1. Introduction

Reanalysis products are typically used for many vary-
ing applications in the earth science community due to
the lack of globally and temporally complete direct ob-
servations (Qian et al. 2006). Examples of the uses of
reanalysis products are to drive land surface models,
datasets. While they are the best approximation of the state of the atmosphere based on both data and dynamic models, the various reanalysis products by the different centers have been found to have many deficiencies on various time scales, especially at smaller (i.e., regional) spatial scales.

For example, a spurious peak in the diurnal cycle of precipitation exists over the Amazon basin in the 40-yr ECMWF Re-Analysis (ERA-40) (Betts and Jakob 2002), and the seasonal cycle of precipitation over this region is too low in both ERA-40 and the ECMWF Interim Re-Analysis (ERA-Interim). Also, the seasonal cycle of 2-m air temperature is higher than observations over the Mississippi River basin for both ERA-40 and ERA-Interim (Betts et al. 2008). The NCEP—National Center for Atmospheric Research (NCAR) reanalysis contains significant biases in the moisture fields in the tropics (Trenberth and Guillemot 1998). Additionally, it is substantially different regionally from the Global Precipitation Climatology Project (GPCP) precipitation data, and it has a diurnal cycle that is much too large over land during the warm seasons (Janowiak et al. 1998). Similarly, Berg et al. (2003) found that the NCEP—NCAR reanalysis and ERA-40 have substantial biases in 2-m air and dewpoint temperatures, surface radiative fluxes, and precipitation over land in North America as compared with various gridded datasets. Aside from the atmospheric fields used to force land surface models, evapotranspiration during the warm season over the U.S. Great Plains region has also been found to be substantially biased in both the NCEP—NCAR reanalysis and ERA-40 (Ruiz-Barradas and Nigam 2005). Previous work has shown that precipitation fields in the newer reanalysis are better than in the older datasets, although this is not universally true for all regions of the globe. For example, ERA-40 corresponds more to the data over the Northern Hemisphere land than it does in the tropics (Bosilovich et al. 2008). Also, 5-day forecast scores of 500-mb height fields show that the CFSR is considerably more accurate than the first generation of the NCEP—NCAR reanalysis (Saha et al. 2010). Similarly, Uppala et al. (2005) found that ERA-40 is enhanced as compared with the first generation of reanalysis products NCEP—NCAR reanalysis and the 15-yr ECMWF Re-Analysis (ERA-15) (Gibson et al. 1997).

Acknowledging these deficiencies, different techniques, such as bias correction (Qian et al. 2006; Berg et al. 2003), and directly combining observations and reanalysis together (Fan and Van den Dool 2004) have been used to derive better land surface forcing datasets. However, the question remains as to the accuracy of the new reanalysis from GMAO [the Modern-Era Retrospective Analysis for Research and Applications (MERRA)], NCEP [the Climate Forecast System Reanalysis (CFSR)], and ECMWF (ERA-40 or ERA-Interim) as compared with the direct observations near the land surface. While research shows that reanalysis products are improving, it is important to evaluate these products directly with near-surface observations over various climate regimes and regions. Specifically, the atmospheric quantities used to force land surface models together with the fluxes of energy and moisture from the land surface to the atmosphere in the various reanalyses must be objectively evaluated.

This study utilizes flux tower observations from 33 different locations in the Northern Hemisphere to evaluate the various reanalysis products from the different centers. The flux network (FLUXNET) is chosen because it provides a large independent body of observations. Such critical evaluations are helpful to researchers who will use these reanalysis products for land–atmosphere interaction studies. They will also help guide the efforts in combining various reanalysis products and in situ data to obtain the best available near-surface atmospheric forcing data for land surface modeling.

2. Datasets

a. In situ observations

The in situ measurements for this study include 33 flux tower locations taken from the flux network of observations (http://www.fluxnet.ornl.gov; Baldocchi et al. 2001). FLUXNET is a global network of micrometeorology measurements of the exchange of water, energy, and CO₂ using eddy covariance techniques at more than 100 towers of differing heights. The data are available through the FLUXNET website. For the present study, the air temperature, wind speed, downward solar radiation, net total radiation, latent heat flux, and sensible heat flux were analyzed for 33 different locations. The locations were chosen so that each has a record of at least 3 yr during the period when reanalysis data are also available. Decker and Zeng (2009) used 17 of these locations to analyze the performance of the Community Land Model and are used again here. Table 1 specifies the locations, period of record used, vegetation type, the tower height, the principal investigator (PI), the contact information for the PIs, and the references for the sites used here. Utilizing numerous towers from across the continent allows for analysis from different vegetation types and climate regimes, including tundra, grassland, and evergreen forests. No gap-filling techniques were utilized on the flux tower observations, but the level-3 data used have quality flags that were used to selectively remove any data that had been deemed of questionable quality by the PI. Even though the questionable quality data are removed from the analysis, flux tower measurements are known to contain errors for various reasons.
<table>
<thead>
<tr>
<th>Name</th>
<th>Abbr.</th>
<th>Lat (°N)</th>
<th>Lon (°W)</th>
<th>Period</th>
<th>Vegetation</th>
<th>Z (m)</th>
<th>PI</th>
<th>E-mail</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGP Main</td>
<td>ARM</td>
<td>36.61</td>
<td>97.46</td>
<td>2003–06</td>
<td>Crop</td>
<td>60</td>
<td>M. Fischer</td>
<td><a href="mailto:mlfischer@lbl.gov">mlfischer@lbl.gov</a></td>
<td>Fischer (2005)</td>
</tr>
<tr>
<td>Atqasuk</td>
<td>Atq</td>
<td>70.47</td>
<td>157.4</td>
<td>2003–06</td>
<td>Wetland</td>
<td>2</td>
<td>W. Oechel</td>
<td><a href="mailto:oechel@sunstroke.sdsu.edu">oechel@sunstroke.sdsu.edu</a></td>
<td></td>
</tr>
<tr>
<td>Audobert Ranch</td>
<td>Aus</td>
<td>31.59</td>
<td>110.5</td>
<td>2002–06</td>
<td>Grassland</td>
<td>4</td>
<td>T. Meyers</td>
<td><a href="mailto:tilden.meyers@noaa.gov">tilden.meyers@noaa.gov</a></td>
<td>Goldstein et al. (2000)</td>
</tr>
<tr>
<td>Blandford Forest</td>
<td>Blo</td>
<td>38.9</td>
<td>120.6</td>
<td>1997–2006</td>
<td>Evergreen needleleaf</td>
<td>12.5</td>
<td>A. Goldstein</td>
<td><a href="mailto:agh@berkeley.edu">agh@berkeley.edu</a></td>
<td></td>
</tr>
<tr>
<td>Bondville</td>
<td>Bo1</td>
<td>40.01</td>
<td>88.29</td>
<td>1996–2006</td>
<td>Croplands</td>
<td>10</td>
<td>T. Meyers</td>
<td><a href="mailto:tilden.meyers@noaa.gov">tilden.meyers@noaa.gov</a></td>
<td>Hollinger et al. (2005)</td>
</tr>
<tr>
<td>Donaldson</td>
<td>SP3</td>
<td>29.75</td>
<td>82.16</td>
<td>1999–2004</td>
<td>Evergreen needleleaf</td>
<td>24.3</td>
<td>T. Martin</td>
<td><a href="mailto:ta.martin@ufl.edu">ta.martin@ufl.edu</a></td>
<td>Gholz and Clark (2002)</td>
</tr>
<tr>
<td>Fermi Prairie</td>
<td>IB2</td>
<td>41.84</td>
<td>88.24</td>
<td>2004–07</td>
<td>Grassland</td>
<td>3.76</td>
<td>R. Matamala</td>
<td><a href="mailto:matamala@anl.gov">matamala@anl.gov</a></td>
<td></td>
</tr>
<tr>
<td>Fort Peck</td>
<td>Fpe</td>
<td>48.31</td>
<td>105.1</td>
<td>2000–06</td>
<td>Grassland</td>
<td>3.5</td>
<td>T. Meyers</td>
<td><a href="mailto:tilden.meyers@noaa.gov">tilden.meyers@noaa.gov</a></td>
<td>Gilmanov et al. (2005)</td>
</tr>
<tr>
<td>Goodwin Creek</td>
<td>Goo</td>
<td>34.25</td>
<td>89.87</td>
<td>2002–06</td>
<td>Grassland</td>
<td>4</td>
<td>T. Meyers</td>
<td><a href="mailto:tilden.meyers@noaa.gov">tilden.meyers@noaa.gov</a></td>
<td></td>
</tr>
<tr>
<td>Harvard Forest</td>
<td>Ha1</td>
<td>42.54</td>
<td>72.17</td>
<td>1991–2006</td>
<td>Deciduous broadleaf</td>
<td>30</td>
<td>J. W. Munger</td>
<td><a href="mailto:jwmunger@seas.harvard.edu">jwmunger@seas.harvard.edu</a></td>
<td>Urbanski et al. (2007)</td>
</tr>
<tr>
<td>Howland Main</td>
<td>Ho1</td>
<td>45.2</td>
<td>68.74</td>
<td>1996–2002</td>
<td>Evergreen needleleaf</td>
<td>60</td>
<td>D. Hollinger</td>
<td><a href="mailto:davidh@hypatia.unh.edu">davidh@hypatia.unh.edu</a></td>
<td>Hollinger et al. (1999, 2004)</td>
</tr>
<tr>
<td>Howland West</td>
<td>Ho2</td>
<td>45.21</td>
<td>68.75</td>
<td>1999–2004</td>
<td>Evergreen needleleaf</td>
<td>60</td>
<td>D. Hollinger</td>
<td><a href="mailto:davidh@hypatia.unh.edu">davidh@hypatia.unh.edu</a></td>
<td>Hollinger et al. (1999, 2004)</td>
</tr>
<tr>
<td>Ivotuk</td>
<td>Ivof</td>
<td>68.49</td>
<td>155.8</td>
<td>2003–06</td>
<td>Open shrublands</td>
<td>3.5</td>
<td>W. Oechel</td>
<td><a href="mailto:oechel@sunstroke.sdsu.edu">oechel@sunstroke.sdsu.edu</a></td>
<td></td>
</tr>
<tr>
<td>Kennedy Space</td>
<td>KS2</td>
<td>28.61</td>
<td>80.67</td>
<td>2000–06</td>
<td>Closed shrublands</td>
<td>3.8</td>
<td>B. Drake</td>
<td><a href="mailto:drakeb@st.edu">drakeb@st.edu</a></td>
<td>Dore et al. (2003)</td>
</tr>
<tr>
<td>Little Prospect Hill</td>
<td>LPH</td>
<td>42.54</td>
<td>72.18</td>
<td>2002–05</td>
<td>Mixed forest</td>
<td>22</td>
<td>J. Hadley</td>
<td><a href="mailto:jhadley@fas.harvard.edu">jhadley@fas.harvard.edu</a></td>
<td></td>
</tr>
<tr>
<td>Lost Creek</td>
<td>Los</td>
<td>46.08</td>
<td>89.98</td>
<td>2000–05</td>
<td>Shrubland wetland</td>
<td>10.2</td>
<td>A. Desai</td>
<td><a href="mailto:desai@os.wisc.edu">desai@os.wisc.edu</a></td>
<td>Sultan et al. (2009)</td>
</tr>
<tr>
<td>Metolius Old Pine</td>
<td>Me4</td>
<td>44.5</td>
<td>121.6</td>
<td>1996–2000</td>
<td>Evergreen needleleaf</td>
<td>47</td>
<td>B. Law</td>
<td><a href="mailto:bev.law@oregonstate.edu">bev.law@oregonstate.edu</a></td>
<td>Law et al. (2003)</td>
</tr>
<tr>
<td>Mize</td>
<td>SP2</td>
<td>29.76</td>
<td>82.24</td>
<td>1998–2004</td>
<td>Evergreen needleleaf</td>
<td>48</td>
<td>T. Martin</td>
<td><a href="mailto:ta.martin@ufl.edu">ta.martin@ufl.edu</a></td>
<td>Gholz and Clark (2002)</td>
</tr>
<tr>
<td>Morgan Monroe</td>
<td>MMS</td>
<td>39.32</td>
<td>86.41</td>
<td>1999–2005</td>
<td>Deciduous broadleaf</td>
<td>18.5</td>
<td>D. Dragoni</td>
<td><a href="mailto:dragoni@indiana.edu">dragoni@indiana.edu</a></td>
<td>Schmid et al. (2000)</td>
</tr>
<tr>
<td>Sky Oaks Old</td>
<td>SO2</td>
<td>33.37</td>
<td>116.6</td>
<td>1997–2006</td>
<td>Closed shrublands</td>
<td>23</td>
<td>W. Oechel</td>
<td><a href="mailto:oechel@sunstroke.sdsu.edu">oechel@sunstroke.sdsu.edu</a></td>
<td>Sims et al. (2006)</td>
</tr>
<tr>
<td>Sky Oaks Young</td>
<td>SO3</td>
<td>33.38</td>
<td>116.6</td>
<td>1997–2006</td>
<td>Closed shrublands</td>
<td>5.8</td>
<td>W. Oechel</td>
<td><a href="mailto:oechel@sunstroke.sdsu.edu">oechel@sunstroke.sdsu.edu</a></td>
<td>Lipson et al. (2005)</td>
</tr>
<tr>
<td>Tonzi Ranch</td>
<td>Ton</td>
<td>38.43</td>
<td>121.0</td>
<td>2001–06</td>
<td>Woody savannas</td>
<td>1.5</td>
<td>D. Baldocchi</td>
<td><a href="mailto:baldocchi@berkeley.edu">baldocchi@berkeley.edu</a></td>
<td>Xu and Baldocchi (2004)</td>
</tr>
<tr>
<td>UCI 1850</td>
<td>CA1</td>
<td>55.88</td>
<td>98.48</td>
<td>2002–2005</td>
<td>Evergreen needleleaf</td>
<td>24</td>
<td>M. Goulden</td>
<td><a href="mailto:mgoulden@uci.edu">mgoulden@uci.edu</a></td>
<td>Goulden et al. (2006)</td>
</tr>
<tr>
<td>UCI 1930</td>
<td>CA2</td>
<td>55.91</td>
<td>98.52</td>
<td>2001–05</td>
<td>Evergreen needleleaf</td>
<td>20</td>
<td>M. Goulden</td>
<td><a href="mailto:mgoulden@uci.edu">mgoulden@uci.edu</a></td>
<td>Goulden et al. (2006)</td>
</tr>
<tr>
<td>UCI 1964</td>
<td>CA3</td>
<td>55.91</td>
<td>98.38</td>
<td>2001–05</td>
<td>Woody savannas</td>
<td>10</td>
<td>M. Goulden</td>
<td><a href="mailto:mgoulden@uci.edu">mgoulden@uci.edu</a></td>
<td>Goulden et al. (2006)</td>
</tr>
<tr>
<td>UCI 1964 Wet</td>
<td>CA4</td>
<td>55.91</td>
<td>98.38</td>
<td>2002–03</td>
<td>Woody savannas</td>
<td>10</td>
<td>M. Goulden</td>
<td><a href="mailto:mgoulden@uci.edu">mgoulden@uci.edu</a></td>
<td>Goulden et al. (2006)</td>
</tr>
<tr>
<td>UCI 1981</td>
<td>CA5</td>
<td>55.86</td>
<td>98.49</td>
<td>2001–05</td>
<td>Evergreen needleleaf</td>
<td>9</td>
<td>M. Goulden</td>
<td><a href="mailto:mgoulden@uci.edu">mgoulden@uci.edu</a></td>
<td>Goulden et al. (2006)</td>
</tr>
<tr>
<td>UCI 1989</td>
<td>CA6</td>
<td>55.92</td>
<td>98.96</td>
<td>2001–05</td>
<td>Evergreen needleleaf</td>
<td>6</td>
<td>M. Goulden</td>
<td><a href="mailto:mgoulden@uci.edu">mgoulden@uci.edu</a></td>
<td>Goulden et al. (2006)</td>
</tr>
<tr>
<td>UCI 1998</td>
<td>CA7</td>
<td>56.64</td>
<td>98.95</td>
<td>2002–05</td>
<td>Evergreen needleleaf</td>
<td>6</td>
<td>M. Goulden</td>
<td><a href="mailto:mgoulden@uci.edu">mgoulden@uci.edu</a></td>
<td>Goulden et al. (2006)</td>
</tr>
<tr>
<td>UMBS</td>
<td>UMB</td>
<td>45.56</td>
<td>84.71</td>
<td>1999–2006</td>
<td>Deciduous broadleaf</td>
<td>10</td>
<td>P. Curtis</td>
<td><a href="mailto:curtis75@osu.edu">curtis75@osu.edu</a></td>
<td>Schmid et al. (2003)</td>
</tr>
<tr>
<td>Vaira Ranch</td>
<td>Var</td>
<td>38.41</td>
<td>121.0</td>
<td>2000–06</td>
<td>Grassland</td>
<td>9</td>
<td>D. Baldocchi</td>
<td><a href="mailto:baldocchi@berkeley.edu">baldocchi@berkeley.edu</a></td>
<td>Xu and Baldocchi (2004)</td>
</tr>
<tr>
<td>Walker Branch</td>
<td>WBW</td>
<td>35.96</td>
<td>84.29</td>
<td>1995–99</td>
<td>Deciduous broadleaf</td>
<td>6</td>
<td>T. Meyers</td>
<td><a href="mailto:tilden.meyers@noaa.gov">tilden.meyers@noaa.gov</a></td>
<td></td>
</tr>
<tr>
<td>Wind River Crane</td>
<td>Wsc</td>
<td>45.82</td>
<td>122.0</td>
<td>1998–2006</td>
<td>Evergreen needleleaf</td>
<td>6</td>
<td>K. Bible</td>
<td><a href="mailto:kbible@uw.edu">kbible@uw.edu</a></td>
<td>Falk et al. (2008)</td>
</tr>
</tbody>
</table>
The flux covariance methods utilized in these measurements do not account for all of the mechanisms by which turbulent energy is transported in the atmosphere. Horizontal advection, subsidence, and low-frequency secondary circulations, such as coherent rolls and convective cells, all result in observational errors (Foken et al. 2011). The magnitude of these errors depends on many factors, including the homogeneity of the measurement site and the tower height. Large-eddy simulation experiments demonstrate that the effect of secondary flows on flux tower measurements increases with increasing tower height (Steinfeld et al. 2007). Because of the above-mentioned difficulties most FLUXNET sites do not actually have a closed energy balance (Aubinet et al. 2000; Wilson et al. 2002). Aside from the above-mentioned causes, the energy balance is also negatively affected by direct measurement errors, such as drift in the shortwave radiation measurements due to the sensors becoming dirty. To access the energy balance closure problem for the locations used in this study, the mean, the mean of the absolute, and the standard deviation of the residual flux (defined as the net radiation minus the sum of the ground, latent, and sensible heat fluxes) are calculated using the half-hourly data. For instance, all of the locations with the exception of Aus, Fpe, MMS, and CA3 have a positive residual energy ranging from 3.46 W m$^{-2}$ for SO3 to 42.17 W m$^{-2}$ for Atq, meaning that the sum of the latent and sensible heat fluxes is on average smaller than the net radiation minus the ground heat flux.

Aside from errors due to the lack of energy balance closure, the different sites in this study use different measuring techniques for other quantities considered here. Only two of the sites use a snow adaptor on the tipping-bucket rain gauge to improve the measurement of snow. Wind speed is also measured differently depending on the site, as some use cup–vane anemometers, while others, such as Me4, use sonic anemometers. It is important to note that, while the flux tower measurements are used in this study to gauge the performance and errors of reanalysis products, the flux tower measurements inherently contain errors as well.

In this study the half-hourly flux data (turbulent, precipitation, and radiative fluxes) were averaged to hourly data to match the time resolution of MERRA and CFSR data, while the observations of the atmospheric states (temperature and wind speed) were filtered by only using every other data point (as the state variables from the reanalysis are not averaged either) to create hourly time series.

b. Reanalysis products

A brief summary of the reanalysis products utilized in this study follows. The important commonalities and differences between the products are highlighted. All of the reanalysis products here utilize conventional data sources (such as wind speed, temperature, moisture, and surface pressure from radiosondes), while other sources, such as precipitation and satellite data, are utilized differently (or not at all) depending on the reanalysis. Only the major observational differences are shown here; for a complete list of the methods and for all of the observations assimilated into each product, consult the references.

1) MERRA

The MERRA product (Bosilovich et al. 2008) is a second-generation reanalysis project from NASA spanning only the satellite era, from 1979 to present, although only the years 1991–2006 are used in the present study. MERRA uses the current Goddard Earth Observing System Data Assimilation System (GEOS-5) (Rienecker et al. 2008). The dynamical atmospheric model in the system is the GEOS-5 AGCM (with observed sea surface temperatures), which includes a finite-volume dynamical core (Lin 2004) with 72 vertical levels and a native latitude–longitude horizontal resolution of only $\frac{1}{2}^\circ \times \frac{1}{3}^\circ$. The land surface model in GEOS-5 is a catchment-based model (Koster et al. 2000) that divides the land surface into various catchments as opposed to regular latitude–longitude grids. The assimilation of observations in GEOS-5 occurs through the use of a three-dimensional variational data assimilation (3DVar) called the NCEP unified gridpoint statistical interpolation analysis scheme (Rienecker et al. 2008). In the analysis cycle, MERRA implements an incremental analysis update (IAU) from Bloom et al. (1996) to slowly update the model. Incrementally applying the update eliminates shocks to the model system by applying frequent small updates as opposed to infrequent large updates. Satellite radiances are directly assimilated while accounting for variational bias correction through the use of the Joint Center for Satellite Data Assimilation (JCSDA) community radiative transfer model. Unlike other reanalysis products, MERRA also assimilates instantaneous rain rate (precipitation) observations though the use of Special Sensor Microwave Imager (SSM/I) and the Tropical Rainfall Measuring Mission (TRMM) Microwave Imager (TMI) measurements.

A key difference between MERRA and other products is the spatial and temporal resolution of the archived data, as MERRA archives much of the model output at its native spatial resolution for every hour. The two-dimensional data, including meteorological states of the first model level, the state of the land surface (including temperature, soil moisture, and canopy-intercepted moisture), turbulent and radiative fluxes, and the components needed to derive the near-surface turbulent fluxes, such as roughness lengths, friction velocities, and stability...
parameters, are all available at hourly time resolutions at the native resolution of $\frac{1}{2}^\circ \times \frac{1}{3}^\circ$. Also, vertically integrated atmospheric quantities, such as precipitable water, tendency terms for water and ice by the different physics packages, including sublimation, evaporation, enthalpy fluxes, and the generation of turbulent kinetic energy, are also stored at this high resolution. The three-dimensional atmospheric quantities, including states and fluxes, are available at a 3-hourly time resolution. The near-surface quantities used here (summarized in Table 2) are all archived at an hourly time scale. MERRA provides several ways to access the data, including a website (http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings).

2) NCEP–NCAR

The NCEP–NCAR reanalysis (NCEP) is a frequently used dataset that covers the period from 1948 to 2009 (the period 1991–2006 is used here) and is the oldest product included in this study (available at http://dss.ucar.edu/pub/reanalysis/). The dynamical atmospheric model used is the NCEP global atmospheric spectral model (operational in 1995) with 28 vertical levels and a horizontal resolution of approximately 210 km (Kalnay et al. 1996). The land surface model is the Oregon State University (OSU) land surface model with two vertical layers. As with MERRA, a 3DVar approach is used to assimilate the observations into the model. The observations include, but are not limited to, rawinsondes, satellite retrievals of atmospheric temperature, satellite-based winds from cloud tracking, aircraft observations of temperature and wind, and land and oceanic surface pressure measurements. While MERRA directly assimilates radiances, the satellite measurements in NCEP are first used to derive vertical temperature profiles before being incorporated into the model (Kalnay et al. 1996). The data are archived at 6-h intervals, with fluxes being archived as the average value over 6 h, while the wind speed and temperature are archived as the instantaneous values.

3) ERA-40

The ERA-40 dataset (available from http://dss.ucar.edu/pub/era40) covers the period from September 1957 to August 2002 (Uppala et al. 2005), with 1991–2001 utilized in the present study. ERA-40 uses the ECMWF dynamical model version cycle 23R4 (CY23R4), with 60 vertical levels at a horizontal resolution of T159 ($\sim$125 km). The data assimilation system used a 3DVar approach with an analysis cycle of 6 h (for both the atmosphere and the land surface). It includes types of data that are very similar to those used by MERRA and other reanalyses but with one significant distinction: ERA-40 includes observed 2-m air temperature (and relative humidity) from weather stations during the six-hourly analysis cycle of the land surface. While the 2-m air temperature is not used directly in the full atmospheric model (the land surface analysis is decoupled from the atmospheric analysis during the assimilation cycle), it indirectly influences the full model by altering the modeled soil temperature and moisture. Also, the observed 2-m air temperatures were utilized in the postprocessing during the output of 2-m air temperatures (Saha et al. 2010). The near-surface atmospheric states and fluxes are archived at 6-hourly intervals.

4) ERA-INTERIM

ERA-Interim (Simmons et al. 2006) is the newest ECMWF reanalysis and covers the period from 1989 to
present (only 2002 is used in the present work). An updated version of the ECMWF forecasting model [version cycle 31r1 (CY31r1)] is used with a horizontal resolution of T213 (~80 km). Aside from the increased resolution over ERA-40, ERA-Interim incorporates full four-dimensional variational data assimilation (4DVar) and is the only reanalysis product in this study to do so (Simmons et al. 2006). While the 3DVar assimilates all observations that occur within a specific period at a single time, the 4DVar is a temporal extension of 3DVar. The cost function used to minimize the difference between the observations and the model is calculated over the 12-h assimilation window. As with ERA-40, the land surface analysis incorporates 2-m air temperature and relative humidity to improve the land surface model temperature and soil moisture fields. ERA-Interim outputs the near-surface fluxes (latent and sensible heat, precipitation, and radiative fluxes) at 3-h intervals as the average over the period. The near-surface wind speed (at 10 m) and air temperature fluxes (latent and sensible heat, precipitation, and radiative fluxes) at 3-h intervals as the average over the period. The near-surface wind speed (at 10 m) and air temperature (at 2 m) are archived at 6-h intervals.

5) CFSR

The CFSR is the newest dataset from NCEP and currently covers from 1979 to the present, and the period 2001/02 is utilized for this analysis. The atmospheric model used in the analysis is the Global Forecast System (GFS) model at a horizontal resolution of T382 (~38 km) with 64 vertical layers (Saha et al. 2010). The land surface component is the Noah land surface model with four vertical layers (Ek et al. 2003). While all of the other reanalyses here use observed sea surface temperatures to drive the atmospheric model, CFSR includes a fully coupled ocean model, the Geophysical Fluid Dynamics Laboratory Modular Ocean Model (GFDL MOM) version 4 (Griffies et al. 2004). The atmospheric analysis occurs at 6-h intervals and uses the gridpoint statistical interpolation (GSI) technique like MERRA. The oceanic assimilation cycle is separate from the atmospheric analysis but also occurs on the same 6-h cycle. The land surface assimilation only occurs once every 24 h and uses observed precipitation. The Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP) 5-day mean (2.5° × 2.5°) gridded and the daily CPC gauge analysis (0.5° × 0.5°) precipitation estimates are combined to drive the land surface model directly to eliminate the biases associated with using model-estimated precipitation rates. Satellite radiances are directly assimilated through the use of radiative transfer models in the atmospheric analysis, and satellite-derived ocean surface wind speeds are used in the oceanic analysis. Unlike the ERA products, the observed 2-m air temperatures are not assimilated. The 10-m wind speed, 2-m surface and first model level temperatures, and the turbulent heat and radiative flux data are archived at hourly intervals, which include both the analyzed states every 6 h but also the model forecasts of the states between these analyses. However, this study does not utilize the hourly quantities because of the discontinuities found in the 6-hourly assimilation cycle (see section 5d). To negate using the hourly forecasts between the 6-hourly analysis cycles, only the 6-hourly CFSR data were utilized in the comparisons.

6) GLDAS

The Global Land Data Assimilation System (GLDAS) produces several different datasets (http://disc.sci.gsfc.nasa.gov/hydrology/data-holdings) that available for use in weather and climate research (Rodell et al. 2004). GLDAS utilizes various different land models from NASA, the National Oceanic and Atmospheric Administration (NOAA), NCAR, and Princeton University–University of Washington to characterize the state of the land surface. The atmospheric data utilized in the creation of the GLDAS forcing dataset varies depending on the period of the data. The GLDAS data utilized in this study have atmospheric states from the NOAA GDAS analysis fields. While the atmospheric states utilize reanalysis products, the precipitation and radiation fluxes are both observation based. The NOAA CMAP data are used for the precipitation field, and the downward shortwave radiation is derived using the Air Force Weather Agency’s Agricultural Meteorological Modeling System (AGRMET) (Fang and Won 2009). The horizontal longitude–latitude resolution of the data used here is 1.0° × 1.0°. The GLDAS forcing dataset consists of 3-hourly air temperature (at 2 m), wind speed (at 10 m), precipitation, and downward solar radiation. Unlike the other reanalysis products in which near-surface atmospheric states and surface fluxes are produced at the same time, surface turbulent and upward radiative fluxes are simulated using four different land models based on the above-mentioned atmospheric forcing in GLDAS. It goes beyond the scope of this study to evaluate surface fluxes from these four land models driven by the same atmospheric forcing data.

3. Methods

a. Vertical interpolation

The tower measurements for each location occur at varying heights. To compare the measured temperature and wind speed at the different heights to the reanalysis output (at either the first model level or a given height of 10 or 2 m), the reanalysis data are interpolated (except for ERA-40, ERA-Interim, and GLDAS, which are discussed later) to the height of the instrument at each location. The method (formulation) that each reanalysis
uses in the model to determine surface fluxes and near-surface air temperatures and wind speeds was used to perform the vertical interpolation.

For MERRA, the air temperature and wind speed were interpolated from the surface utilizing the bulk transfer formulations found in GEOS-5 (Rienecker et al. 2008) to the level of the tower observations. The formulations relate the surface Richardson number to the friction velocity and temperature scale. All the necessary quantities for interpolating the data are included in the archive (friction velocity, temperature scale, Richardson number, and roughness lengths for momentum and heat), so that the interpolation is direct and noniterative.

To interpolate the NCEP–NCAR reanalysis, the atmospheric states were interpolated between the first model level and the surface using the Monin–Obukhov similarity functions from the NCEP-Noah land surface model used in the NCEP global atmospheric spectral model. Because the NCEP–NCAR reanalysis only archives friction velocities as 6-hourly averages and atmospheric states as instantaneous values, it is inappropriate to use the averaged friction velocity for interpolation. To circumvent this issue, the temperature at two heights (the surface and 2 m) and the 10-m wind speed were used to iteratively solve for the friction velocities and the stability parameter. These values, together with the archived roughness length, were then used to interpolate the data.

The same procedure was used to interpolate the CFSR air temperatures and wind speeds to the tower heights, as both NCEP–NCAR and CFSR use variations of the OSU (Noah) land surface model. The air temperatures from the surface and the first model level, along with the wind speed from the first model level and the roughness length, are iteratively used to calculate the values for the stability parameter and the friction velocities. These were then used to interpolate the temperature and wind speed from the ground to the tower height.

The air temperature and wind speed were not interpolated to the tower heights for both the ERA-40 and ERA-Interim data. When archiving near-surface temperature and wind speed, ERA-40 and ERA-Interim utilize a modified Monin–Obukhov scheme to interpolate the data to produce temperatures and wind speed that correspond well to those observed by surface weather stations. ERA-40 uses a large (much larger than other models, including ERA-Interim) roughness length to account for the drag that orography induces on the atmosphere. While providing the needed drag, the use of such large roughness lengths prevents the interpolation from the land surface to 2- or 10-m height directly. Therefore, ERA-40 uses an empirically formulated blend of winds and temperatures from a 75-m height to derive near-surface quantities (White 2003). Although ERA-Interim uses smaller roughness lengths that do not include orographic drag, the near-surface wind and temperature are still empirically found in the same way as they are in ERA-40. Because of the care and expertise in deriving these methods to interpolate the data, we feel that any attempt to improve on these interpolations would result in less valid values. The negligible impact of vertical interpolation on the analysis is discussed in section 5c, affirming that not interpolating the ERA datasets does not degrade the analysis.

The GLDAS forcing dataset was not interpolated because the dataset only includes air temperatures and wind speeds at a single level, and also does not include roughness lengths or friction velocities. Since not enough data are included to interpolate the atmospheric states, they were left unaltered.

b. Temporal averaging and interpolation

The atmospheric states and fluxes were evaluated on different time scales. First, the mean annual cycle is evaluated by computing monthly values averaged over the entire period of observations for each station. The average and standard deviation were then calculated for each dataset and location based on the mean monthly values. Finally, for each site the bias, correlation, root-mean-square error (RMSE), and the standard deviation of the errors were calculated between the tower observations and the reanalysis products.

The 6-hourly data are also analyzed in addition to the monthly-mean quantities. For the fluxes of energy and water (i.e., latent and sensible heat fluxes, radiative fluxes, and precipitation), each dataset is temporally averaged from its native resolution (hourly for MERRA, three hourly for ERA-Interim and GLDAS) to six hourly. The NCEP, CFSR, and ERA-40 fluxes already have a 6-hourly temporal resolution and do not require any processing. The state variables, however, are output as instantaneous values and therefore are not averaged over the 6-h period. Rather, the hourly data for MERRA and CFSR (or 3 hourly for GLDAS and ERA-Interim) values that correspond to the 6-hourly values from NCEP and ERA-40 are simply utilized without averaging. The correlation, bias, ratio of standard deviations, standard deviation of the errors, and root-mean-square error were then found by comparing the tower observations to the reanalysis from specific months.

The mean diurnal cycle for various months is also presented. The mean diurnal cycle is found (at a 6-hourly temporal resolution) by first averaging all of the days within a month to a single mean diurnal cycle for that month. Then, the statistics (mentioned previously) are calculated between the observations and the reanalysis products using the mean diurnal cycles for several different months.
The diurnal cycle from June to August (JJA) and from December to February (DJF) are considered here.

To facilitate a comparison for all of the sites in a single figure, histograms of the correlation, the ratio of the standard deviation of reanalysis to tower observations, bias, and root-mean-square error are plotted separately for the statistics found by comparing the monthly data, 6-hourly data, and mean diurnal cycles.

c. Decomposition of the mean square error

The mean square error (MSE) was calculated for every variable and site at the three different time scales. To quantify the contribution of the bias (accuracy), correlation (the correspondence), and difference in standard deviation (the agreement in the amount of variation), the mean square error was decomposed (Gupta et al. 2009) as

\[
MSE = 2\sigma_d\sigma_{obs}(1 - \rho) + (\sigma_d - \sigma_{obs})^2 + (\mu_d - \mu_{obs})^2,
\]

where \(\rho\) is the correlation between the reanalysis and the tower observations, \(\sigma_d\) is the standard deviation of the reanalysis, \(\sigma_{obs}\) is the standard deviation of the observations, \(\mu_d\) is the mean of the reanalysis, and \(\mu_{obs}\) is the mean of the tower observations. The first term is the contribution of the correlation to MSE, the second term is the contribution of the differences in standard deviation between the observations and the reanalysis, and the third term is the bias contribution. Dividing (1) through by MSE, the three terms become the relative contribution to MSE from the three different statistics. While the relative contribution does not aid in defining which of the reanalysis is better than another, it allows for an easy interpretation of the reasons for the disagreement between the reanalyses and the measurements.

d. Ranking

Because of the large number of sites and quantities that are presently compared, a ranking system is utilized to determine which dataset has the best performance for a given quantity. The ranking system similar to Brunke et al. (2003) is chosen because of its simplicity. Because of the different periods for the different reanalysis products, only stations with data for all of the products are compared. Since different locations have nonmissing data for different periods depending on the variable, the number of stations utilized to calculate the average rank varies depending on the variable. For a given variable (e.g., temperature) at a given station (e.g., Blo in Table 1), the bias is calculated for each product using the 6-hourly data. Then, each station is given a ranking score (for temperature between 1 and 6; for sensible and latent heat fluxes between 1 and 5, since GLDAS turbulent fluxes from offline modeling using four different land models are not evaluated), with 1 given to the product with the lowest absolute bias and 6 (or 5) given to the product with the highest absolute bias. The procedure is repeated for each station and then averaged over all of the stations. For example, a product with the lowest absolute bias at every single station would have a bias ranking score of 1. The result is the average ranking over all of the towers for each product. The same method is also utilized for the standard deviation of the errors.

4. Results

a. Temperature

The histogram illustrating the correlation, bias, ratio of standard deviations, and root-mean-square error for 6-hourly air temperature is shown in Fig. 1. All of the reanalysis products have a correlation of greater than 95% for nearly every single station in terms of the 6-hourly air temperature. The ratio of standard deviations also shows that the products match the observed 6-hourly variation in temperature very well, as nearly all of the stations fall within a range of 0.9–1.1, meaning that there is a less than 10% error for the variation of temperature. However, CFSR tends to underestimate the variability in the 6-hourly temperature, as seven stations (Aud, Fpe, KS2, LPH, Los, SP2, and Wsc) have a ratio of standard deviations of less than 0.9. NCEP clearly underestimates the temperature much more often than it overestimates the temperature, while other reanalyses all tend to overestimate the 6-hourly temperature (i.e., have a positive bias) approximately as often as they underestimate the temperature. In fact, 10 locations for NCEP have a bias of less than −4.5 K. The root-mean-square error for all of the products is greater than 1 K. Yet, it is clear from Fig. 1 that NCEP has the worst-performing air temperature, as 17 locations have an RMSE of greater than 2.5 K, while all other products have more than half of the stations with an error less than 2.5 K.

The rankings calculated for both the bias and the standard deviation of the errors are shown in Table 2. The 6-hourly temperature bias is lowest for ERA-Interim (2.71), followed closely by MERRA (2.86) and GLDAS (2.79). However, while the temperature variability is again the best for ERA-Interim (1.43), MERRA, GLDAS, CFSR, and ERA-40 are much closer together and distinctly not as good as ERA-Interim, even though the ratio of standard deviations falls between 0.9 and 1.1 for all of the datasets at most of the stations (Fig. 1).

As stated previously, the monthly-mean diurnal cycle for different months was also analyzed. Figure 2 shows the histogram from the analysis for the mean diurnal
cycles of temperature in June–August. For ERA-40 and ERA-interim, nearly all of the locations have a correlation of greater than 0.95, whereas CFSR, MERRA, and NCEP have several stations with a correlation between 0.9 and 0.95. MERRA, ERA-40, and ERA-Interim have a variability (ratio of standard deviations) between 0.9 and 1.1 for 13, 12, and 18 locations, respectively. NCEP, however, tends to overestimate the variability, with 12 stations having a ratio of standard deviations between 1.1 and 1.3. CFSR and GLDAS both underestimate the variability at the greatest number of locations (9, including the same 7 from Fig. 1 and ARM and Ivo for CFSR, and 10 locations for GLDAS), unlike the other reanalysis products. However CFSR also overestimates the variability at 10 locations. It is also clear from Fig. 3 (and from Fig. 1) that NCEP has a much larger RMSE than the other reanalyses under these circumstances, as 12 stations have an RMSE greater than 3 K. ERA-40 and ERA-Interim have the largest number of stations, with an RMSE of less than 1 K (7 and 11, respectively), while MERRA and CFSR have 18 and 15 locations, respectively, with an RMSE of less than 2 K.

The mean diurnal cycle of temperature in December–February has a much weaker cycle because of the decreased incoming solar radiation as compared with June–August. While ERA-40, ERA-Interim, and GLDAS have the vast majority of stations with a correlation of greater than 0.95, MERRA has a nearly equal number of stations with a correlation between 0.9 and 0.95 as it does between 0.95 and 1.0, and NCEP has only slightly more stations with a correlation greater than 0.95 as it does between 0.9 and 0.95. Clearly, NCEP, MERRA, ERA-40, and CFSR are better at simulating the strong diurnal cycle of temperature in the summer than the weaker cycle in the winter. For the winter, the spread of the ratio of standard deviations is much larger for all of the reanalyses, and CFSR has a particularly degraded diurnal variability in the winter. The bias in ERA-40 is much larger in DJF than during JJA, and this is reflected in the large number of stations with a RMSE of greater than 3 K. The stations with large negative biases are similar among most of the reanalysis datasets. MERRA, NCEP, CFSR, and GLDAS all have a bias of less than −5 K for both Blo and Var. ERA-40, however, has the smallest bias at Blo at only −0.82 K, while ERA-Interim is the only reanalysis with a positive bias at Blo, with a value of nearly 1.7 K. The cause for the large warm bias at Blo for ERA-Interim as opposed to the other products is that the afternoon temperature in ERA-Interim is much higher here than in both the observations and the
other reanalysis products even though the nighttime temperature is more accurate. In terms of RMSE, ERA-Interim clearly does the best for the wintertime diurnal cycle, although GLDAS also does well.

The relative contributions of the correlation and the bias to the total mean squared error [from (1)] for the six-hourly, daily, and monthly air temperatures from each station are shown in Fig. 4. On a monthly time scale, the bias is the largest contributing factor to the mean squared error for all of the datasets, as most of the locations have a bias term greater than 0.5 (meaning more than half of the MSE is due to the bias). As the averaging period becomes shorter, however, the bias becomes much less of a contributing factor. For the daily temperatures, the bias term is still consistently greater than 0.5 for most of the stations. However, the 6-hourly data show that the bias term does not contribute as significantly to the MSE. In general, as the averaging period becomes shorter, Fig. 4 shows that the correlation term becomes the dominant cause of the MSE.

b. Wind speed

Although the mean monthly temperature was very well correlated with the observations for all of the reanalysis products, the monthly wind speed has a correlation greater than 0.95 for only a handful of stations (Fig. 5). MERRA has a correlation between 0.4 and 0.8 for eight stations, while only two stations (Bo1 and MMS) have a correlation greater than 0.95. NCEP, ERA-40, CFSR, and GLDAS have a clear peak in the correlation histogram, with the greatest number of stations having a correlation between 0.4 and 0.7, much lower than that for temperature. It is also clear from Fig. 5 that all of the reanalysis products except GLDAS have a strong tendency to overestimate the variability in the wind, with most of the stations having a ratio of standard deviations between 1.7 and 5.0. The reanalysis that best simulates the variability in wind speed is GLDAS, as eight stations fall between 0.7 and 0.9. The bias is positive for nearly every station for all the different reanalyses. The RMSE is quite low for the majority of the products except for a couple of locations for MERRA, NCEP, ERA-Interim, and CFSR. Both MERRA and CFSR have an RMSE greater than 3.0 m s\(^{-1}\) at Wsc, while ERA-40, GLDAS, and NCEP all have an RMSE of approximately 1.8 m s\(^{-1}\) there.

The 6-hourly wind speed analysis yields very similar results to that from the monthly analysis, with a few notable exceptions. The correlation between the products and the observations is lower for the 6-hourly data, and the spread in the variability histogram and the RMSE are both slightly greater. The rankings in Table 2 again show that ERA-Interim has the lowest bias on average, with a ranking of 2.57, while GLDAS is only slightly larger, with a score of 2.93. The bias ranking is the worst for MERRA (4.36), although CFSR and NCEP are both
close at 4.14 and 3.93, respectively. The rankings based on the standard deviation of errors again show ERA-Interim to be the best, with a score of only 1.50, and show other datasets scoring much higher, with MERRA, ERA-40, and GLDAS scoring 3.21, 3.43, and 3.93, respectively. However, NCEP and CFSR clearly have the worst rankings (5.36 and 4.43, respectively) for the standard deviation of the errors.

The analysis of the average diurnal cycle for the months of June–August as well as December–February (figures not shown) provides similar conclusions as that of the monthly averages and the 6-hourly data. The correlation for all of the reanalysis products is very poor as compared with that of temperature. The diurnal cycle in wind speed was very weak in the tower observations for nearly every location (the amplitude was typically $-1\,\text{m s}^{-1}$), and thus it seems reasonable to not expect the reanalyses to be able to match such a weak diurnal cycle. However, MERRA and ERA-Interim have the lowest values of root-mean-square error and bias as compared with other datasets.

The contributions of MSE for wind speed behave in a manner similar to those for temperature; however, certain differences are evident. For the monthly averages, Fig. 6 shows that the bias is the dominant factor in MSE, as a majority of the stations are grouped near the point (1, 0). However, for the 6-hourly data, all of the reanalyses show that both the correlation and the bias are important terms, with the level of the importance depending on the location. The daily data, in contrast, behaves like either the 6-hourly or the monthly data, depending on the reanalysis product. For MERRA, the daily wind speed MSE is composed of nearly entirely the bias term, with only six stations having a bias contribution term less than 0.7. This result is not surprising, given the poor bias ranking for 6-hourly wind speeds for MERRA. Other products, however, have 10 or more locations with a bias contribution term of less than 0.7 on a daily time scale. The 6-hourly data behave similarly to the daily data, as MERRA is the only dataset with a significant number of stations where the bias contribution term is greater than 0.6. The standard deviation term, however, is much less important than both of these terms, as the majority of the points in Fig. 6 follow very close to the line $x + y = 1$ (with $x$ and $y$ being the normalized bias and correlation terms, respectively).

c. Precipitation

Similar to the analysis of the monthly wind speed, the correlation of monthly precipitation rate is much poorer than that of temperature, as shown in Fig. 7. For MERRA, almost a third of the stations have a monthly precipitation correlation of greater than 0.9; however, a nearly equal number have a correlation between 0.4 and 0.7. CFSR performs very similar to MERRA in terms
of precipitation, as is seen in Fig. 7. Of all the reanalyses, NCEP overestimates the variability in monthly precipitation the most, as approximately a third of the stations have a ratio of modeled standard deviation to observed between 1.1 and 1.7. ERA-40 greatly underestimates the variability in precipitation at five locations, as the ratio of standard deviations is less than 0.7. MERRA does a very good job of capturing the variability in precipitation, as half of the stations have a ratio of 0.7–1.3. The biases in monthly precipitation are the lowest for MERRA, CFSR, and ERA-40. While ERA-40 does a very poor job of capturing the monthly variability in precipitation, it does a good job in capturing the mean amount. The root-mean-square error shows that GLDAS has the best overall monthly precipitation as compared with the other reanalyses, as nearly all of the stations have a root-mean-square error of less than 0.5 mm day$^{-1}$. This is not surprising because GLDAS precipitation is from observed CMAP gridded precipitation data. The location that is the most poorly simulated among the different products is Wsc, as NCEP, ERA-40, ERA-Interim, CFSR, and GLDAS all have an RMSE greater than 2.5 mm day$^{-1}$ there. However, Blo and LPH have an RMSE greater than 2.5 mm day$^{-1}$ for NCEP, CFSR, and GLDAS, while MERRA, ERA-40, and ERA-Interim all simulate the monthly precipitation well at these two locations.

The 6-hourly precipitation, like the monthly precipitation, is poorly correlated among all of the datasets, as shown in Fig. 8, especially for SO2. MERRA, NCEP, ERA-Interim, CFSR, and GLDAS all have a correlation of less than 0.2 at SO2. For GLDAS, 50% of the stations have a correlation greater than 0.5, while CFSR, ERA-Interim, ERA-40, NCEP, and MERRA have 56%, 50%, 32%, 39%, and 50% of the locations, respectively, with a correlation greater than 0.5. The variability is greatly underestimated for all of the reanalyses. In particular, it is less than 20% of the observed variability at SO2. MERRA, CFSR, and NCEP tend to have positive biases much more so than negative biases. However, all of the products have a bias less than $-1.5$ mm day$^{-1}$ at SO2.

Fig. 4. The relative contributions to the MSE from the correlation and bias terms in (1) at each station plotted against each other for each of the six reanalysis products using the 6-hourly, daily, and monthly-mean temperature. The MSE contribution from the standard deviation term in (1) is not explicitly plotted, although the monthly-mean temperature error contribution terms are grouped close to (1, 0). This means that the majority of the MSE on this time scale corresponds to the bias term in (1), while the 6-hourly and daily data are farther from this point, indicating that bias contributes less (but still significantly) to the MSE on these time scales.
The rankings from the precipitation rate (Table 2) show that overall, GLDAS is the least biased product (2.14) because of its use of observed gridded precipitation, followed by ERA-40, ERA-Interim, and MERRA with scores of 3.0, 3.5, and 3.71, respectively. CFSR and NCEP are clearly more biased overall, with scores of 4.21 and 4.43, respectively. The variability of the precipitation errors in Table 2 again shows that GLDAS has the best-performing precipitation (2.14); however, ERA-Interim also performs quite well (2.29). Unlike the bias ranking, MERRA (3.36) performs better than ERA-40 (4.29). NCEP is again the worst among all of the products.

The 6-hourly precipitation is also analyzed as a function of season by comparing DJF and JJA separately (figures not shown). The correlation of 6-hourly precipitation is higher among all the datasets in DJF. MERRA, NCEP, ERA-40, ERA-Interim, CFSR, and GLDAS have 9, 5, 3, 9, 8, and 4 stations, respectively, with a correlation of greater than 0.8 for DJF; they only have 1, 0, 2, 0, 0, and 0 stations, respectively, with a correlation of greater than 0.8 in JJA. Although the correlation is much higher in DJF than JJA, there is a much larger positive bias in DJF for all of the data except for GLDAS. Similar to the positive bias, all of the data have a much larger ratio of standard deviations in DJF as opposed to JJA. These discrepancies between DJF and JJA are partly caused by both the instrumentation of the flux towers and the spatial/temporal variability of precipitation as discussed in section 5b.

The monthly averaged diurnal cycle of precipitation, shown in Fig. 9, is again poorly correlated between all of the datasets and the tower observations, with a correlation of less than 0.7 over the vast majority of stations. The correlation for all of the reanalyses is equally poor for both JJA and DJF (figure not shown). However, GLDAS has 10 stations that have a correlation greater than 0.7. The bias is the lowest for MERRA, ERA-40, and GLDAS, as there are 13, 12, and 15 stations, respectively, with a bias between −0.5 and 0.5 mm day$^{-1}$. ERA-Interim, CFSR, and NCEP, however, have 14, 11, and 11 stations, respectively, with a bias between 0.5 and 1.5 mm day$^{-1}$. The RMSE shows that overall, GLDAS has the best diurnal cycle of precipitation, as 11 stations have an RMSE $<1$ mm day$^{-1}$, confirming the results from Table 2. All of the reanalyses, except CFSR, generally have an RMSE $<2$ mm day$^{-1}$. The three stations (Goo, Ha1, and LPH) where MERRA has an RMSE $>3.5$ mm day$^{-1}$ are the stations where CFSR has an RMSE $>4$ mm day$^{-1}$. However, the majority of the errors in subdiurnal precipitation is not related to the bias.

The relative contribution to the MSE from the three terms in (1) (Fig. 10) is very different for precipitation on the 6-hourly time scale than it is for temperature or wind speed. While the bias term contributes more than...
40% of the MSE for the monthly data, the bias term is almost zero for every reanalysis for the 6-hourly and the daily data. The implication of this result is that the rankings from the standard deviation of the errors is more significant than those from the biases because the biases contribute so little to the total error on 6-hourly time scales. The correlation component is the largest term on short time scales for nearly every station for MERRA, NCEP, ERA-40, ERA-Interim, and CFSR. Only GLDAS has a significant number of stations (9) with a correlation term of less than 0.5, again because of its use of observed gridded precipitation data. Unlike wind speed and air temperature, the major source of error in all of the datasets in the precipitation at subdiurnal and daily time scales is the timing of the events.

d. Latent heat flux

The monthly averaged latent heat flux is well correlated for all of the reanalyses at many locations (Fig. 11). Nearly half of the stations have a correlation of greater than 0.85 for MERRA, while a majority of the stations for ERA-40 and ERA-Interim have a correlation of greater than 0.9. While MERRA is well correlated, it tends to overestimate the monthly variability, as the ratio of standard deviations is frequently 1.7–2.5. ERA-Interim also tends to overestimate the variability, while ERA-40 has 10 stations with a ratio between 1.1 and 1.5. For MERRA, the bias is less than 25 W m\(^{-2}\) at 7 stations, and 8 stations have a bias between 25 and 50 W m\(^{-2}\). CFSR, ERA-Interim, and ERA-40 have approximately two-thirds of the stations with monthly averaged biases of less than 25 W m\(^{-2}\). NCEP, in contrast, has two-thirds of the locations with a bias greater than 25 W m\(^{-2}\). All of the reanalyses have a tendency to overestimate the latent heat flux (i.e., a positive bias). However, at Blo, SO2, and SO3, all of the reanalyses (except for ERA-Interim at SO3) underestimate the latent heat fluxes by more than 10 W m\(^{-2}\) on average annually, and both NCEP and CFSR have an RMSE greater than 30 W m\(^{-2}\) at Blo. The root-mean-square error is less than 40 W m\(^{-2}\) at 11 stations for MERRA, 19 for ERA-40, 12 for ERA-Interim, and 13 for CFSR. The stations with the largest RMSE are the locations that also coincide with the largest biases as well.

The histogram of statistics from the mean diurnal cycle over all months for the latent heat flux is shown in Fig. 12. Since the diurnal cycle of surface turbulent fluxes is driven by the available energy at the surface, it is not surprising that MERRA, NCEP, ERA-40, and ERA-Interim all have a correlation of greater than 0.9 for nearly every location (see section 4f). The stations that
have relatively poor correspondence ($\rho < 0.9$) compared with the observed diurnal cycles of latent heat fluxes for the aforementioned datasets are Blo, Me4, SO2, SO3 (all except ERA-Interim), and Var. While MERRA has 10 stations that have $\rho$ between 0.9 and 0.95 and 11 that have $\rho$ greater than 0.95, it also has only four stations (listed above) with $\rho < 0.4$. Therefore, the only stations that MERRA has a poor correspondence with the observations are the stations that none of the reanalysis products are able to mimic the observations. Of these four locations, three are from immature and young forests (SO2, SO3, and Me4) with canopy heights less than 10 m, which may partially explain why all of the reanalyses perform poorly at these locations. CFSR has a very poor correlation on the diurnal time scale (as 14 stations have $\rho < 0.9$, while only 12 have $\rho > 0.9$). The diurnal variability of latent heat flux is much too large in all of the reanalysis datasets because of the overestimation of the late afternoon latent heat flux. All of the reanalyses overestimated the 6-hourly variability as well. ERA-Interim, MERRA, and ERA-40 also have a small bias in the latent heat flux at a majority of the stations, with ERA-Interim having 18 stations with a bias between $-10$ and $10$ W m$^{-2}$. CFSR overestimates the mean latent heat flux at nearly half (13) of the stations with a bias between 10 and 25 W m$^{-2}$.

The 6-hourly latent heat flux is much less correlated and has a larger RMSE than the diurnally averaged latent heat flux for all of the reanalysis products (Fig. 13). While less correlated among all of the datasets, only NCEP has an appreciable number of locations (10 out of 19) that has a correlation less than 0.7. MERRA is poorly correlated at several locations (Blo, SO2, and SO3); however, at these locations, the other reanalysis products are also poorly correlated. Both ERA-40 and ERA-Interim have correlations less than 0.7 at SO2 and SO3; CFSR is poor at SO2 and Blo. It is clear from Fig. 13 that MERRA, NCEP, and ERA-Interim tend to overestimate the variability in the 6-hourly latent heat flux. Also, all of the products have a tendency to overestimate the mean amount of latent heat, as the bias is positive for the majority of the locations, and NCEP almost universally has a bias from 25 to 50 W m$^{-2}$.

The rankings from the 6-hourly latent heat fluxes (Table 2) show that ERA-40 is clearly the least biased product with a score of 1.93, and CFSR is also less biased overall than the other products with a score of 2.73. The standard deviation of errors rankings again shows that ERA-40 has the best 6-hourly latent heat flux (2.13), and ERA-Interim has the second best variability with a score of 2.4.

The bias contributes significantly to the MSE of latent heat flux on monthly time scales but much less significantly on 6-hourly and daily time scales (figure not shown), as the contributions to MSE behave much like
those for wind speed. All of the reanalyses show that the bias term is greater than 0.5 for nearly every location for the monthly data. However, on the 6-hourly time scale, the correlation term is dominant, as it is less than 0.5 only for 6, 9, 3, 5, and 4 stations for MERRA, NCEP, ERA-40, ERA-Interim, and CFSR, respectively. Similar to air temperature and wind speed, the standard deviation does not contribute significantly to the MSE.

e. Sensible heat flux

The monthly sensible heat flux is again well correlated to many of the observations for all of the reanalysis datasets except for NCEP (Fig. 14). While MERRA, ERA-Interim, and CFSR have 48%, 56%, and 66% of the total stations with a value of ρ greater than 0.9, respectively, 52% of the stations for NCEP have ρ < 0.7. Clearly, ERA-40 and NCEP have a worse representation of the monthly sensible heat flux than MERRA, ERA-Interim, and CFSR. Like the latent heat flux, the reanalysis products tend to overestimate the variability in sensible heat flux, as MERRA, NCEP, and CFSR have 5, 9, and 8 stations, respectively, with a ratio of standard deviations greater than 1.5 (i.e., the variability is more than 50% more than it should be). ERA-40 tends to underestimate the variability as much as it overestimates it, as six locations have a ratio of standard deviations greater than 1.3 while five locations have a ratio less than 0.7. At three of the five locations that MERRA greatly overestimates the variability (Blo, Los, and SP2), NCEP and CFSR also overestimate the variability there (with the ratio of standard deviation > 1.7). Of these three stations that are simulated so poorly, Blo and Los are also overestimated (ratio > 1.5) for ERA-Interim, which has the most accurate monthly variability, as nine stations have a value between 0.9 and 1.1. While MERRA overestimates the variability at several stations, including the stations that no reanalysis simulates well, it also has the lowest bias among the products. For MERRA, 53% of the stations have a bias between –10 and 10 W m⁻², while for ERA-40 (ERA-Interim) the fraction is 39% (42%). ERA-Interim also tends to underestimate the bias by more than 10 W m⁻² (at 58% of the locations). The RMSE is lowest for ERA-Interim, as three stations have an RMSE of less than 10 W m⁻². MERRA has 13 stations with an RMSE < 30 W m⁻², while ERA-40 has 17 stations and ERA-Interim has 16 stations.

The histogram based on the 6-hourly sensible heat fluxes (Fig. 15), much like Fig. 13, shows that MERRA, ERA-Interim, and ERA-40 are the least biased among the datasets. However, the variability in ERA-Interim most closely matches the observations for the largest number of stations, while NCEP and CFSR tend to greatly overestimate the variability quite frequently. CFSR has a ratio of greater than 1.5 at Aud, Blo, Fpe,
KS2, Los, Ton, and Var, while NCEP is the largest at Aud, Blo, Los, Wsc, and Var. Similar to NCEP and CFSR, MERRA has a ratio greater than 1.5 at both Blo and Var. The biases in sensible heat fluxes are nearly evenly distributed in MERRA around the $-10$ to $10$ W m$^{-2}$ bin, while CFSR has a large positive bias ($>25$ W m$^{-2}$) at Aud, Ton, and Var (three of the seven locations where the variability is greatly overestimated). NCEP also has the largest positive bias ($>25$ W m$^{-2}$) at Aud. The RMSE follows the results discussed above, as NCEP and CFSR have an RMSE $>75$ W m$^{-2}$ for the majority of the locations. ERA-Interim is the only dataset with a significant number of stations (6) with an RMSE $<40$ W m$^{-2}$, while MERRA has two and ERA-40 has four. Interestingly, Aud and Var are two of the locations where the RMSE for ERA-Interim is less than 40 W m$^{-2}$.

The rankings from the 6-hourly bias and standard deviation of errors (Table 2) clearly summarize the results from Fig. 15. The bias is lowest for ERA-40 (2.27), followed closely by MERRA (2.53) and ERA-Interim (2.80). The standard deviation of errors rankings, however, demonstrate that ERA-Interim has a better variability (1.80), followed closely by ERA-40 (2.27) and MERRA (2.60). From Table 2, it is clear that ERA-40, ERA-Interim, and MERRA have 6-hourly sensible heat fluxes more like the observations than both CFSR and NCEP.

The monthly averaged diurnal cycle sensible heat flux (figure not shown) is simulated very accurately by all of the reanalysis products, with the exception of NCEP. As is the case with the diurnal cycle of latent heat flux, ERA-Interim, ERA-40, and MERRA are all very well correlated with the observations, an expected finding since the diurnal cycle of sensible heat flux is driven primarily by the net radiative fluxes.

The contributions to the MSE for the sensible heat flux follow a very similar pattern to that for the latent heat flux on both monthly and 6-hourly time scales (figure not shown). The errors for all of the reanalyses on a monthly time scale are dominated by the bias term, while on 6-hourly and daily scales the errors are dominated by the correlation term, with a smaller contribution from the variability term.

f. Incoming shortwave radiation

The mean monthly incoming shortwave radiation is perhaps the best simulated quantity compared in this study among all of the reanalysis datasets (Fig. 16). Nearly
all of the stations have a correlation greater than 0.95 and a ratio of the standard deviations between 0.9 and 1.3 for all of the products. However, MERRA, ERA-Interim, and NCEP all have more stations with a ratio between 1.1 and 1.3 than they do between 0.9 and 1.1. Therefore, these datasets tend to overestimate the monthly variability of incoming shortwave radiation as compared with the observations. Interestingly, the bias for MERRA and NCEP is nearly always positive, and neither of these datasets has a bias less than $-10 \text{ W m}^{-2}$. While ERA-40, GLDAS, and CFSR all have at least one station with a bias less than $-10 \text{ W m}^{-2}$ (Los, CA3, and CA6 for ERA-40, and Aud for both GLDAS and CFSR), they too tend to overestimate the incoming shortwave radiation. NCEP is the only reanalysis with a significant (24) number of stations with a bias greater than $25 \text{ W m}^{-2}$.

The diurnal cycle of downward solar radiation (figure not shown) at the surface is nearly perfectly correlated with the observations for all the reanalysis products. The variability of diurnal shortwave radiation is the best for ERA-40, ERA-Interim, CFSR, and GLDAS, as 95%, 81%, 62%, and 60%, respectively, of the stations have a ratio of standard deviations between 0.9 and 1.1. NCEP, however, overestimates the diurnal variability in shortwave radiation, as 78% of the locations have a ratio greater than 1.1. While MERRA slightly overestimates the variability with the ratio between 1.1 and 1.3 for 48% of the stations, it does so much less than NCEP. The reason that both MERRA (and to a much greater extent NCEP) overestimates the variability is due to the peak values of shortwave radiation in the late afternoon being too large. These large values are reflected in the bias, as a third of the stations have a bias between 10 and 20 W m$^{-2}$ for MERRA and nearly half of the stations have a bias between 40 and 60 W m$^{-2}$ for NCEP. The RMSE for ERA-40, ERA-Interim, and CFSR are

---

**Fig. 10.** As in Fig. 4, but using precipitation rate (mm day$^{-1}$).
considerably lower at a majority of the locations than for MERRA and NCEP.

The 6-hourly incoming shortwave radiation (figure not shown) is again very well correlated among all of the products; however, ERA-Interim is the only product to have the majority of the locations with a correlation greater than 0.95, as the other five products typically have a correlation between 0.9 and 0.95. The variability is between 0.9 and 1.1 for every location for MERRA, and nearly every location for CFSR, GLDAS, and the ECMWF products. Only NCEP and GLDAS overestimate the variability in 6-hourly downward shortwave radiation (with a ratio greater than 1.1) for an appreciable number of locations (SP3, Fpe, Los, SP2, and Wsc for NCEP; and SP3, Fpe, Los, MMS, and SP2 for GLDAS). The bias tends to be small and positive among all of the datasets; however, NCEP is the only reanalysis where the bias is nearly always greater than 25 W m\(^{-2}\). The bias and standard deviation of the errors rankings in Table 2 again show that ERA-Interim has the smallest overall bias and standard deviation of errors. For the bias ranking, NCEP is clearly the worst among all of the products, as the score of six means that it has the largest bias at every site. However, GLDAS obviously has the worst standard deviation of errors ranking, even though the bias ranking is better than both MERRA and NCEP.

The MSE from the 6-hourly, daily, and monthly shortwave radiation, like the turbulent fluxes, is again primarily due to the bias term on monthly time scales and the correlation term on 6-hourly and daily time scales (figure not shown). Unlike the turbulent fluxes, NCEP has a bias term that is also significant on 6-hourly time scales. All of the other reanalyses have a correlation contribution greater than 0.8 for nearly every location at 6-hourly time scales. NCEP, however, has a correlation contribution less than 0.8 at 15 locations. At these sites, the remaining MSE is from the bias term, not the standard deviation term.

g. Net surface radiation

As the net surface radiation is comprised of the downward solar (and also the upward shortwave and net longwave) radiation, the monthly net surface radiation, similar to the shortwave radiation, is produced accurately by all of the reanalysis products (figure not shown). While
the variability and timing of monthly net radiation is adequate among all of the products, MERRA is the only dataset with nearly every station having a bias between $-10$ and $10 \, \text{W m}^{-2}$. The low bias for MERRA translates to the lowest RMSE among all of the datasets, with the majority of the stations having an RMSE from $10$ to $15 \, \text{W m}^{-2}$.

The 6-hourly net radiation behaves in a similar manner to the downward shortwave radiation; however, several important differences are evident from Fig. 17. While the shortwave radiation is nearly perfectly correlated with the observations for all of the reanalysis products, the timing of the net radiation is slightly worse. However, MERRA, NCEP, ERA-Interim, and CFSR all have correlations greater than 0.9 at all of the sites. ERA-40, in contrast, has a correlation less than 0.9 at MMS, WBW, and CA6. The variability and biases of net radiation is nearly identical for MERRA and ERA-Interim, as both have a tendency to overestimate both quantities at the same seven locations. From Fig. 17, MERRA and ERA-Interim obviously have lower RMSE values at the most stations, even though the RMSE is typically $50$–$70 \, \text{W m}^{-2}$. The rankings for the 6-hourly net radiation bias (Table 2) show that, surprisingly, NCEP has the lowest bias in the 6-hourly net radiation ($2.46$), followed by MERRA ($2.92$). However, from the ranking of the standard deviation of the errors, NCEP obviously has the worst variability, while MERRA and ERA-Interim (with scores of $2.31$ and $2.38$) are better than the other reanalysis products.

5. Discussion

a. Errors at individual sites

There are several locations where all (or most) of the reanalysis datasets do not simulate well the quantities considered here, and the physical interpretation of the results enables us to identify the probable cause of these errors. All reanalysis products have poor turbulent fluxes over SO2, SO3, and Me4. Also, MERRA has a very large temperature bias at Me4, while the other reanalysis products also have a larger bias and RMSE at Me4 than they do at other stations. For SO2 and SO3, MERRA, NCEP, ERA-40, and CFSR all underestimate the mean

![Fig. 12. As in Fig. 1, but for each of the five reanalysis products considered here for the monthly averaged diurnal cycle of latent heat over all months (W m$^{-2}$).]
precipitation. ERA-40 underestimates the precipitation at SO2 the most (with a bias of $-1.9$ mm day$^{-1}$), while NCEP, GLDAS, CFSR, MERRA, and ERA-Interim have biases of 0.7, 0.58, 0.65, 0.54, and 0.62 mm day$^{-1}$, respectively. The underestimation of latent heat flux due to a lack of available moisture causes the overestimation of sensible heat flux at these locations, which contributes to the near-surface temperatures being too high. For CFSR, the temperature bias is greater than 4 K at SO2, while MERRA, NCEP, ERA-Interim, and ERA-40 all have large RMSEs at these two locations. Another possible cause for the general disagreement between all of the reanalyses and the tower observations at these points could be due to the vegetation, but the poor performance over these locations is not seen in other sites with similar vegetation types. Two of these poorly simulated sites (SO3 and SO2) have closed shrublands, while Me4 is in an evergreen needleleaf forest. The different closed shrubland site KS2 is simulated much better than SO3 and SO2, and therefore the vegetation type alone cannot explain the poor performance across all the reanalysis products.

Another site where the reanalysis products did not simulate well is Blo. The temperature and turbulent fluxes for CFSR, NCEP, ERA-40, and ERA-Interim are all especially poor at Blo. Similar to SO2 and SO3, the root cause of the poor representation of this site is most probably the bias in precipitation. The four aforementioned datasets have a negative bias in precipitation ranging from $-2.16$ for NCEP to $-1.42$ mm day$^{-1}$ for ERA-40. Consequently, CFSR, NCEP, ERA-40, and ERA-Interim all underestimate the mean latent heat flux at Blo by approximately 15 W m$^{-2}$, which is approximately one-third to one-quarter of the error (converting from mm day$^{-1}$ to W m$^{-2}$) in precipitation. MERRA is the only reanalysis to have a (small) positive bias in precipitation at Blo (0.2 mm day$^{-1}$) and consequently has a (small) positive bias in latent heat flux (3.6 W m$^{-2}$). Also, like SO2 and SO3, the smaller latent heat in CFSR, NCEP, ERA-40, and ERA-Interim is partially compensated for by an overestimation of sensible heat flux. The agreement between the observed total turbulent (latent plus sensible) heat flux (discussed below) supports the conclusion that it is the inaccurate
partitioning of the turbulent fluxes at these sites that leads to the poor performance.

Although it is sometimes possible to deduce the causes for the disagreements between the datasets at certain locations, other sites do not have errors that can be traced so easily. At Var, CFSR, NCEP, MERRA, and GLDAS all had a large cold temperature bias there. However, all of the datasets at Var (except GLDAS) had a positive bias in both incoming shortwave and net surface radiation. Understandably, these reanalyses also had large values of RMSE for the turbulent fluxes as well, especially ERA-40 and ERA-Interim, even though the ECMWF products had lower values of temperature biases than the other datasets. Although MERRA has the largest bias in precipitation there (1.26 mm day$^{-1}$)—ERA-40 and ERA-Interim had biases of 0.63 and 0.16 mm day$^{-1}$, respectively—all of these products overestimated the latent heat flux by approximately 1 mm day$^{-1}$ (~30 W m$^{-2}$). A consequence of the positive bias in both downward shortwave and net surface radiation is that all of the products overestimated sensible heat fluxes as well as latent heat fluxes. While reanalysis products performed poorly at certain locations, the majority of the locations had atmospheric states and surface fluxes simulated well by the reanalysis.

**b. Uncertainties due to measurement error and scale differences**

Part of the disagreement between the reanalysis products and tower observations is due to the scale differences. The gridded reanalysis products represent the grid-average values of each variable, with the size of a grid box being as large as a few hundred kilometers. This is especially important for precipitation with its high spatial and temporal variability, and for areas with large spatial variability in topography on subgrid scales. The evaluation of reanalysis products would be better carried out by incorporating multiple nearby point observations in such cases, although our focus here is more on the integrated evaluation across different variables (which is ideally done by flux towers).

Aside from spatial–temporal problems with precipitation, the measurements themselves can be problematic. For example the tipping-bucket method utilized for

---

**Fig. 14.** As in Fig. 1, but for each of the five reanalysis products considered here for the monthly averaged sensible heat flux (W m$^{-2}$).

<table>
<thead>
<tr>
<th></th>
<th>MERRA</th>
<th>NCEP</th>
<th>ERA-40</th>
<th>ERA-int</th>
<th>CFSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_{\text{ratio}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
precipitation is known for introducing errors during snowfall. Only two of the sites (Los and Me4) in this study utilize a rain gauge with a snow adapter, and thus the precipitation results could be sensitive to errors in snowfall measurement. To quantify both the effect of season and the effect of snowfall on the analysis, a comparison of 6-hourly precipitation in JJA and DJF are analyzed as well as precipitation when the temperature is greater than 0°C (although mixed-phase precipitation can exist near the freezing point, a simple cutoff is chosen). During JJA, convective rain events are typically much smaller in scale than synoptic snowfall events of DJF because of the different driving mechanisms of precipitation in the varying seasons. Therefore, it is plausible that the limited spatial scale of precipitation might impact the analysis more in the summer than the winter. For JJA, all of the datasets exhibited a ratio of standard deviations of much less than one for nearly every station, which is indicative of the reanalysis products not producing the magnitude of the convective summer rain events at the tower locations.

For DJF, however, all of the datasets had a large number of locations with the ratio of standard deviations much greater than one. Also, while the biases were small in JJA, MERRA, NCEP, ERA-40, ERA-Interim, and CFSR all had several stations with large positive biases in DJF. The comparison utilizing the precipitation only when the temperature is above the freezing point did not change the results significantly. The biases in 6-hourly precipitation (using all months) are generally positive even when only rainfall is considered, and the correlations for all of the reanalysis products are nearly identical to the results from using all of the precipitation data. Similarly, the RMSE for all of the products remained relatively unchanged and the histogram of RMSE (figure not shown) followed nearly an identical pattern to that using all of the precipitation data. While DJF precipitation had a larger positive bias than JJA for all products except GLDAS, the analysis using only rainfall did not show a reduction in the overall bias. While the bias, correlation, and ratio of standard deviations changed dramatically with the season, they did not significantly change when only considering rainfall. Overall, the RMSE did not vary significantly between DJF, JJA, or when only considering rainfall for any of the reanalysis products.
the reanalysis products. Therefore, it is concluded that while errors in snowfall measurement are present when using tipping-bucket methods, they did not significantly affect the conclusions of this study.

To further investigate the possible errors caused by errors in the flux tower measurements, the total turbulent heat flux (the sum of the latent and sensible heat fluxes) is also analyzed. The total turbulent heat flux at the land surface is controlled by the net available radiation, as energy balance dictates that the net radiation minus the ground heat flux be partitioned between the latent and sensible heat fluxes. For all of the reanalysis data, it is found that the 6-hourly total turbulent heat flux is in much better agreement with the measurements than the latent or sensible heat fluxes separately. The correlation (figure not shown) for the 6-hourly total turbulent heat flux is greater than 0.9 for nearly every station for all of the reanalysis products. Similar to the net radiation, the ratio of standard deviations show a distribution of about one for all products even though the latent heat flux has a large number of stations with a ratio much greater than one for both MERRA and ERA-40. While the biases for sensible and latent heat fluxes range from −25 to 50 W m$^{-2}$ for most locations for all of the products, the biases in the total turbulent heat fluxes are significantly smaller, only ranging from −15 to 15 W m$^{-2}$ for nearly every location and dataset. Similar to the net radiation, MERRA and ERA-Interim have a positive bias in the total turbulent heat flux at 7 and 9 locations, respectively, while NCEP, ERA-40, and CFSR have 7, 10, and 9 stations, respectively with a positive bias. The RMSE for the latent-plus-sensible heat fluxes is significantly lower than the RMSE for the sensible or latent heat flux, with values typically around 10 W m$^{-2}$. Clearly, the reanalysis products and the flux towers agree well in terms of the total turbulent heat fluxes. The positive bias for the total turbulent heat fluxes can be qualitatively attributed to the energy closure balance problem in flux tower measurements discussed in section 2a. However, the bias of the total turbulent flux from the flux tower measurements is much smaller than both the latent and sensible heat fluxes; therefore, the large discrepancies seen in sections 4d and 4e are the result of how the reanalysis products partition the turbulent fluxes and not due to the energy closure problem.

c. Impact of vertical interpolation

While specific sites can have the errors of different quantities traced to each other, the differences between the reanalysis products can be partially attributed (qualitatively) to the different methods used by the various
reanalysis products. Although ERA-40 and ERA-Interim are not interpolated in this study, these two datasets are accurate in terms of the air temperature. Clearly, not interpolating these two datasets did not hinder the comparison of air temperature from them. The empirical methods used at ECMWF to output both near-surface air temperature and wind speed (by combining the near-surface value and the value at 75 m above the surface, depending on the roughness length) provide good estimates of the actual near-surface temperature and wind speed. Also, the impact of interpolating the temperature for MERRA, NCEP, and CFSR is small, yet overall it results in a lower RMSE for the majority of the locations. The RMSE based on the interpolated temperatures from MERRA is lower by approximately 0.08 K for half of the locations versus using the MERRA 2-m temperatures. Similarly, CFSR also shows that 50% of the stations have an RMSE at least 0.05 K lower than when the 2-m temperatures are used directly. NCEP shows the largest improvement when the temperature is interpolated, as the RMSE is reduced by greater than 0.18 K at half of the stations. While interpolating the temperature results in a lower RMSE for more than half of the stations, the magnitude of the improvement is small and not all locations show improvement. These results confirm that while interpolating the temperature to the tower height can...
slightly improve the results, the lack of interpolating ERA-40, ERA-Interim, and GLDAS do not significantly degrade the analysis.

The wind speed, however, shows much larger improvements from the interpolation. For MERRA and CFSR, interpolating the wind speed improves the RMSE by greater than 0.5 and 0.2 m s⁻¹, respectively, for three-quarters of the locations. NCEP, however, shows a much smaller improvement when the wind speed is interpolated. The enhanced improvement of interpolated wind speed versus the interpolated temperature is expected, because the wind speed is zero near surface and increases with height in the surface layer. In contrast, the surface temperature usually does not deviate much from air temperature over vegetated areas.

d. Qualitative impact of the differences between the reanalysis products

The accuracy of the ECMWF products is most certainly partially attributable to the use of 2-m air temperature in the output of these quantities at ECMWF. While MERRA, NCEP, and CFSR do not utilize 2-m air temperatures during the assimilation procedure, the monthly air temperatures are generally very reasonable. Also, the bias (summarized in Table 2) for the 6-hourly air temperature in MERRA is nearly as good as ERA-Interim even though MERRA does not assimilate the 2-m air temperature observations. While the temperature bias is low for both ERA-Interim and MERRA, ERA-Interim clearly has the lowest standard deviation of errors for all of the quantities considered here except for the turbulent fluxes.

The use of observed precipitation in GLDAS and the land data assimilation system in CFSR is evident in the analysis. While CFSR’s performance is similar to MERRA and ERA-40 in terms of the standard deviation of errors ranking for temperature, it has a lower bias overall for latent heat than MERRA does. The primary factor for the improved simulation of latent heat in CFSR may be related to the use of observed precipitation in the land surface data assimilation cycle in CFSR. Through the use of observed precipitation, CFSR produces accurate mean latent heat fluxes even though the temperature (and consequently, the sensible heat fluxes) is more biased than MERRA and ERA-Interim. The CMAP precipitation in GLDAS is the reason that GLDAS has the lowest precipitation bias and standard deviation of errors among the products. However, because of the different scales between the CMAP observations and the in situ observations used here, differences still exist between GLDAS and tower precipitation fields.

Even though CFSR and MERRA utilize essentially the same observations in the assimilation process, major differences exist between the two datasets. CFSR has a much better correlation with the observations for 6-h temperature than hourly temperature due to the assimilation increment being applied once every 6 h. The correlation of the hourly CFSR wind speed and temperature data are much less than that of the 6-hourly CFSR data (generally less than 0.95 compared to greater than 0.95 for the mean diurnal cycle of temperature), because of the analysis only occurring every 6 h. In contrast, the hourly MERRA data are as well correlated as the 6-hourly data for temperature. When only using the 6-hourly data, the impact of the analysis cycle in CFSR cannot be seen; however, it becomes apparent in the hourly data. Figure 18 is a time series of air temperature at Var on 10 June 2001 and shows an example of the “jumps” that are sometimes seen in the hourly CFSR data products. The time series in Fig. 14 shows the temperature in CFSR is rising too quickly in the afternoon, and at 1200 UTC (the analysis time) the series suddenly jumps downward only to continue rising again later. The sudden drop in air temperature shown in Fig. 14 is not the result of a physical process, as there is not a precipitation event or sudden drop in incoming solar radiation. However, MERRA does not contain such jumps, as the analysis increment is applied over several time steps between analysis cycles. Therefore, the hourly MERRA data are not hindered by this limitation as the CFSR hourly data are.

e. Relative contributions to MSE

The importance of examining the relative contributions of MSE can be seen by carefully examining the methods currently used to adjust reanalysis products to create land surface forcing datasets. The method employed by Qian et al. (2006) scales the reanalysis using available observations. The reanalysis is scaled such that the monthly mean of the reanalysis will equal the monthly mean of the observations. Utilizing such a technique, the bias is the only statistic examined here that is guaranteed to match the observations. The standard deviation of the reanalysis is not directly scaled to match the observations, even though it changes according to the scaling factor used to alter the bias and the correlation between the two may not change much. In light of this information, it is clear that all of the reanalysis products cannot be greatly improved at the monthly time scale, because a majority of the MSE comes from the bias on this time scale. However, at 6-hourly time scales, none of the reanalysis products can be improved significantly, as the bias is not the largest source of the errors. To improve the reanalysis greatly, a method that accounts not only for monthly bias but also correlation and variance at submonthly time scales would need to be developed in the future. Until then, one may
need caution in the use of reanalysis products at subdaily time scale. For example, our study suggests that land models forced by a global dataset that is based on reanalysis (e.g., Qian et al. 2006) will inherently have difficulties in producing a realistic hourly evolution of surface fluxes.

Similar to the air temperature, the monthly average wind speed has errors partly due to the biases. On a 6-hourly time scale, however, the bias is still significant for all of the reanalyses. Even though the correlation of the diurnal wind speed between all of the reanalyses and the observations is poor, the bias still plays a significant role in the RMSE. Therefore, correcting the 6-hourly (or hourly) wind speed in each of the reanalyses using previously mentioned methods is not only straightforward but very useful.

6. Conclusions

This research evaluated the reanalysis products from GMAO, NCEP, and ECMWF using FluxNet observations from 33 different locations. The quantities considered were the atmospheric variables used to force land surface models and the turbulent fluxes of energy and water near the surface. The sources of the mean square error (MSE) were examined on both 6-hourly and monthly time scales, and the overall performance of each of the reanalysis products for each atmospheric state and land surface flux was examined to quantify the differences, strengths, and weaknesses of the new second-generation reanalysis products.

The near-surface air temperature is best simulated by ERA-Interim, followed closely by MERRA, ERA-40, and GLDAS on 6-hourly, monthly, and monthly averaged diurnal time scales. The bias in the air temperature is lowest for ERA-Interim; however, the bias is also small in MERRA overall. ERA-Interim has a much lower standard deviation of errors for the temperature than the other reanalysis products. Interpolating the temperatures from MERRA, NCEP, and CFSR to the tower heights slightly improves the RMSE for more than half of the locations. When comparing flux tower observations and the reanalysis, the disagreement in heights between the two datasets is not a significant source of the errors. Therefore, it is not surprising that ERA-Interim had the lowest bias in temperature even though the reanalysis data are from a different height than the tower observations.

The RMSE errors in the monthly wind speed are comparable among all of the reanalysis datasets. MERRA, NCEP, ERA-40, CFSR, GLDAS, and to a lesser extent ERA-Interim, all overestimate the monthly variability of wind speed. Unlike temperature, the wind speed is not well correlated on a diurnal time scale with any of the reanalyses. The biases in the 6-hourly wind speed are lowest for ERA-Interim, although not much more so than GLDAS and ERA-40. However, the ranking based on the standard deviation of the errors clearly shows that ERA-Interim is better than the other products.

Overall, GLDAS has the lowest precipitation bias and standard deviation of errors, because it directly uses CMAP observations. MERRA directly assimilates satellite rain rates, yet the resulting reanalysis product does not agree with the tower measurements as well as GLDAS. ERA-Interim, in contrast, does not assimilate any satellite rain rates, yet it performs better than MERRA in terms of the bias and standard deviation of errors.

The monthly turbulent fluxes are well simulated in MERRA, ERA-40, ERA-Interim, and CFSR, and the diurnal cycle of turbulent fluxes is much better correlated with the observations than the wind speed. MERRA has the most accurate sensible heat flux, while ERA-Interim has the best latent heat flux at monthly averaged diurnal time scales. The differences between the tower observations and the reanalysis products cannot be explained by the lack of energy balance closure in the tower measurements. The total turbulent heat flux (latent plus sensible) is in much closer agreement between all of the reanalysis products and the measurements. The slight positive bias in the total turbulent flux seen in the reanalysis can at least be partially attributed to the lack of energy balance closure in the measurements; however, this bias is much smaller in magnitude than the biases in sensible and latent heat fluxes.

All of the reanalyses overestimate the incoming shortwave radiation, with NCEP overestimating the most followed by MERRA. ERA-Interim, again, has the best incoming shortwave radiation.

The mean square errors in monthly reanalysis precipitation can be largely attributed to the bias term in all of the reanalysis. However, the 6-hourly reanalysis products, while still containing biases, have errors primarily caused by the timing of precipitation events. Similarly, the MSE errors for temperature, wind speed, and turbulent fluxes are dominated by the lack of correlation at 6-hourly and daily time scales, while the bias only becomes the dominant source of errors at monthly time scales. Unlike temperature and the turbulent fluxes, however, wind speed still has an appreciable amount of error at 6-hourly time scales associated with the bias. Data constructed from reanalysis for the purposes of forcing land surface models [such as those from Qian et al. (2006) and Berg et al. (2003)] need to address the fact that the MSE errors are caused more by the correlation term than the bias term at daily and 6-hourly time scales. Future efforts to produce a best possible forcing dataset for land models need to not only account for the monthly biases but also address the lack of correlation on time scales shorter than monthly.
Overall, the ERA-Interim dataset appears to be slightly better than all of the rest, at least when qualitatively combing the results from all of the quantities analyzed here. Even though MERRA utilizes a smaller grid than ERA-Interim, the increased spatial resolution does not result in closer agreement with the observations. However, the usefulness of the hourly data from MERRA is apparent because its native temporal resolution is much more similar to that of land surface models.

Acknowledgments. This work was supported by NASA (Grant NNX09A021G), NOAA (Grant NA10NES4400006), and NSF (Grant AGS-0944101). We thank Hoshin Gupta for suggesting the analysis of the mean square error decomposition and two anonymous reviewers for their insightful suggestions (including the impact of energy balance closure issue in tower measurements on reanalysis evaluations). We thank the scientists at NCAR CISL, as they provided web access to the reanalysis products from NCEP and ECMWF. Also, we thank NCAR for the use of the NCAR computers for obtaining data from the mass store system. The GLDAS data used in this study were acquired as part of the mission of NASA’s Earth Science Division and archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC). We thank all of the principal investigators for each of the towers utilized in this study, each contributed their data to the FluxNet database, as their efforts have produced an abundance of valuable data for the earth sciences community. Finally, we thank the Oak Ridge National Laboratory for providing access to the FluxNet data.

REFERENCES


