Selection of a Reanalysis Meteorological Product for Use in Solar Resource Assessment

Anthony Lopez

Follow this and additional works at: https://digitalcommons.du.edu/geog_ms_capstone

Part of the Geographic Information Sciences Commons, and the Spatial Science Commons

Recommended Citation
https://digitalcommons.du.edu/geog_ms_capstone/66

This Capstone is brought to you for free and open access by the Department of Geography and the Environment at Digital Commons @ DU. It has been accepted for inclusion in Geography and the Environment: Graduate Student Capstones by an authorized administrator of Digital Commons @ DU. For more information, please contact jennifer.cox@du.edu,dig-commons@du.edu.
Selection of a Reanalysis Meteorological Product for Use in Solar Resource Assessment

Anthony Lopez

University of Denver Department of Geography

Capstone Project

For

Master of Science in Geographic Information Science

October 25, 2015
Abstract

The U.S. Department of Energy's National Renewable Energy Laboratory and collaborators are updating the National Solar Radiation Database with the next-generation satellite-based solar resource assessment known as the Physical Solar Model (PSM). This is part of an effort to achieve NREL and DOE’s goal to make solar energy technologies cost competitive with other forms of energy in the United States. Required in this effort is the selection of an appropriate meteorological reanalysis product that provides critical inputs for solar energy estimation. This analysis quantitatively and qualitatively assessed the spatial and temporal accuracy of three meteorological reanalysis products using ground measurements. It was found that the NASA MERRA and CFSR reanalysis products had similar accuracies while the NARR reanalysis exhibited the least accuracy. NASA MERRA exhibited superior accuracy in total precipitable water and atmospheric pressure, which made its use in radiative transfer modeling most justifiable.
# Table of Contents

Abstract .................................................................................................................. 2

Acronyms .................................................................................................................. 4

Introduction .............................................................................................................. 4

Problem Statement ................................................................................................. 6

Literature Review for Statistical Methods of Comparison to Ground Measurements ........................................................................................................... 7

Root Mean Square Error (RMSE) ......................................................................... 8

Mean Absolute Error (MAE) .................................................................................. 9

Coefficient of Variation of Root Mean Square Error (CV(RMSE)) .................. 11

Percent Bias ........................................................................................................... 12

Methodology ........................................................................................................... 13

Analysis Extent ..................................................................................................... 15

Ground Measurement Data .................................................................................... 17

Reanalysis Products ............................................................................................... 17

Modern-Era Retrospective Analysis for Research and Applications (MERRA) .... 17

American Regional Reanalysis (NARR) .............................................................. 18

Climate Forecast System Reanalysis (CFSR) ..................................................... 18

Included Variables ................................................................................................. 18

Total Precipitable Water ....................................................................................... 19

Wind Speed .......................................................................................................... 19

Dew Point .............................................................................................................. 19

Surface Air Temperature ....................................................................................... 19

Atmospheric Pressure ......................................................................................... 20

Spatial Lookup ..................................................................................................... 20

Temporal Decimation and Averaging ................................................................... 20

Statistical Methods .............................................................................................. 21

Results Storage .................................................................................................... 22

Challenges ............................................................................................................. 22

Results .................................................................................................................... 23

Total Precipitable Water ....................................................................................... 23

Wind Speed .......................................................................................................... 26

Dew Point .............................................................................................................. 30

Surface Temperature ............................................................................................ 34

Atmospheric Pressure ......................................................................................... 38

Discussion .............................................................................................................. 42

Appendix ............................................................................................................... 44

Analysis Python Code ........................................................................................... 44

References ............................................................................................................. 50
Acronyms

NASA National Aeronautics and Space Administration
MERRA Modern-Era Retrospective Analysis for Research Applications
NARR North American Regional Reanalysis
CFSR Climate Forecast System Reanalysis
ISD Integrated Surface Database
NREL National Renewable Energy Laboratory
NOAA National Oceanic Atmospheric Administration
UW University of Wisconsin
NSRDB National Solar Radiation Database
RMSE Root Mean Square Error
MAE Mean Absolute Error
CVRMSE Coefficient of Variation of Root Mean Square Error
HDF Hierarchical Data Format
DOE United States Department of Energy
**Introduction**

The National Renewable Energy Laboratory (NREL) in conjunction with the University of Wisconsin (UW) and the National Oceanic Atmospheric Administration (NOAA) have been developing a physics based solar irradiance gridded dataset for inclusion in the NREL National Solar Radiation Database (NSRDB). The collaboration intends to produce a freely available dataset for use in energy systems across the western hemisphere at \( \frac{1}{2} \) hour time-steps from 1998 to 2014 with a spatial resolution of 4 x 4 km. Of importance to the solar irradiance calculations themselves (known as radiative transfer models) and with implications for energy system modeling is surface meteorological variables, including: wind speed, temperature, atmospheric pressure, and precipitable water. There are two distinct classes of data products available for such variables, ground measurements and atmospheric or meteorological reanalysis data. Ground measurements refer to commercially available sensors setup at discrete locations. Reanalysis is a term used to describe a systematic approach to produce climate-monitoring data using an unchanging data assimilation scheme (Climate Data). Ground measured surface data of the meteorological variables are available from the Integrated Surface Database (ISD – a NOAA product); however, the data is spatially sparse and would have to undergo spatial interpolation for assignment to the solar grid. As a
result, atmospheric reanalysis datasets are being considered as these datasets are spatially continuous and span long periods of time; however, there are reliability issues that vary depending on time, location, and variable. (Climate Data). There are numerous publications measuring the accuracy of reanalysis datasets using ground measurements; however, a literature review revealed no such analysis focusing on the select meteorological variables and the spatial extent required for inclusion in a solar resource database, such as the NSRDB.

**Problem Statement**

Which atmospheric reanalysis product demonstrates the best relative accuracy—both in space and time—when compared to ground measurements.

**Method**

This analysis statistically evaluated and identified the most appropriate atmospheric reanalysis product for inclusion in the new NSRDB solar grid based on their relative accuracy. The three datasets evaluated were: NASA’s Modern-Era Retrospective Analysis for Research and Applications (MERRA), NOAA’s North American Regional Reanalysis (NARR), and NOAA’s Climate Forecast System Reanalysis (CFSR). While MERRA and CFSR are global datasets NARR only cover’s North America. These datasets were evaluated using several statistical measures—discussed in detail below—against 216 measured ISD
stations located in the contiguous United States. Each statistical method was applied to daily and monthly means spanning years 2005 to 2009. Several of the resulting measures were mapped to reveal spatial trends in the bias of each of the atmospheric reanalysis models. It is shown that each atmospheric reanalysis dataset has its own unique bias both spatially and temporally. It is also shown that there is not a clear demarcation between all atmospheric reanalysis products and their accuracy. Given such, each of the atmospheric reanalysis variables and their associated statistical measures were described both quantitatively and qualitatively. Ultimately, the MERRA reanalysis product was chosen based on a minimization of bias for the total precipitable water variable, a variable with the greatest impact on solar resource estimation from among the chosen variables.

**Literature Review for Statistical Methods of Comparison to Ground Measurements**

A literature review returned numerous publications measuring the accuracy of atmospheric reanalysis datasets using ground measurements; however, no analysis focusing on the select meteorological variables and the spatial extent required for inclusion in the NSRDB solar grid existed. Thus, the following literature review focuses on the methods employed for statistically evaluating the accuracy of modeled data using ground measurements.
**Root Mean Square Error (RMSE)**

The RMSE (equation 1) is a statistical measure typically used in geoscience to compare modeled data with measured data. It represents the sample standard deviation of the differences between predicted and observed values (Root-mean-square Deviation). Chai and Draxler, 2014 describe the advantages of using RMSE for evaluation of model performance as it provides a complete picture of the error distribution. Chai and Draxler, 2014 also describe the advantage of RMSE over Mean Absolute Error (MAE) in that MAE is best at describing uniformly distributed errors, while models generally produce normally distributed errors – something RMSE is better at representing. Jia et al., 2013 rely on RMSE to demonstrate their solar model accuracy over East Asia. They also incorporate scatter plots, Mean Bias Error (MBE), R, and $R^2$ for an effective illustration of their model accuracy (Figure 1).

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}.
\]

(Equation 1)
Figure 1 Source Jia et al., 2008. Correction to Evaluation of ERA-40, NCEP-1, and NCEP-2 reanalysis air temperatures with ground-based measurements in China

**Mean Absolute Error (MAE)**

The MAE (equation 2) is a statistical method used to measure differences between forecasts and outcomes. It is usually used to evaluate the accuracy of continuous variables (Ma et al., 2008) by returning the average of the absolute errors (Mean Absolute Error). Willmott and Matsuura conclude MAE to be an appropriate and "most natural" measure in their report on the applicability of MAE to climate datasets as it is an unambiguous measure of average error magnitude.
Gueymard and Wilcox also use the MAE to assess the overall model accuracy of solar resource from radiometric measurements and predictions from models using ground-based and satellite data (see figure 4). They present their findings on an annual basis per measurement location. Ma et al., 2008 conducted a model accuracy test for atmospheric reanalysis data in China using MAE as a main statistical measure. They used the resulting MAE per station to reveal model bias based and trends for topographic features (figure 2).

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i|
\]

(Equation 2)

Figure 11. Statistically significant trends (95%-level) of 1979–2001 annual air temperature from (a) observations, (b) ERA-40, (c) NCEP-1, and (d) NCEP-2. Red and blue circles denote positive and negative trends, respectively. Larger circles represent larger trends.

Figure 2 Source Ma et al., 2008. Correction to Evaluation of ERA-40, NCEP-1, and NCEP-2 reanalysis air temperatures with ground-based measurements in China.
**Coefficient of Variation of Root Mean Square Error (CV (RMSE))**

The CV (RMSE) (equation 3) normalizes the RMSE by the mean measured value. It is a convenient method for comparing the errors of different variables with different magnitudes. For example, errors from wind speed with a range of 0 to ~14 meters/second would be hard to compare to errors from air temperature in any temperature measure.

Habte et al. used CV (RMSE) in their publication of differences between solar irradiance datasets and demonstrated its effectiveness via output maps (see figure 3). Given the different magnitudes associated with each dataset, it would be difficult (or impossible) to distinguish any meaningful patterns from the maps without first normalizing the RMSE. Boilley and Wald, 2015 utilize CV (RMSE) for an intercomparison of surface solar irradiation between ERA-Interim and MERRA atmospheric reanalysis data products. The authors present CV (RMSE) in tables with multiple locations between the two datasets that allow for easy comparison.

\[
CV(RMSE) = \frac{RMSE}{\bar{y}}
\]

(Equation 3)
Percent Bias

Percent bias is a measure of the average tendency of modeled values to be larger or smaller than measured values \( (R: \text{Percent Bias}) \).

Percent bias can be negative or positive with 0 being optimal or no bias in the modeled values compared to the measured. Negative values indicate underestimation while positive values indicate overestimation. Anderberg et al. use percent bias as an intercomparison between two different modeled solar datasets and a measured solar dataset. When plotted, percent bias can reveal seasonal or annual trends—like the negative bias seen in figure 4. It
can also highlight time periods of extreme bias, also seen in figure 3 in the month of November.

Figure 4 Source Anderberg et al., 2010. Evaluating Solar Resource Variability from Satellite and Ground-Based Observations

**Methodology**

As mentioned previously, the resources chosen contained the required variables for solar resource assessment and have adequate spatial and temporal resolution and extent. It should be noted that the resources had already been collected and stored on NREL analysis servers and had been converted to a standard structure using the Hierarchical Data Format (HDF) – this was from previous efforts at NREL. The analysis was executed in sequential order. First, by defining the analysis extent (s) and ground measurement data, then by defining the atmospheric reanalysis products to include in the analysis. Third was to define the
variables and reasoning to include in the statistical analysis. In the fourth step, a spatial lookup between the ground measurements and the atmospheric reanalysis data was performed. In the fifth step, the data was decimated and averaged to coincident time-steps. In the sixth step, statistical methods were applied to the data. Finally, the data and results were stored in a single HDF intended for graphing and mapping. A concept of the analysis can be seen in figure 5.

* Design applies to other reanalysis products as well
* Design does not represent all combinations of statistics and times

**Figure 5 Analysis Concept flow chart. Author: Anthony Lopez**

All steps are covered in more detail in the following sections. Python (programming language) with a suite of scientific modules, including: Numpy, Scipy, h5py, and Pandas along with IPython Notebook were
used to perform the analysis. The analysis can be reproduced using the Python code found in the appendix.

**Analysis Extent**
The full analysis extent spans the contiguous United States for 216 distinct ISD station locations (figure 6).

![ISD Station](image)

*Figure 6 Full Extent of locations included in the analysis. Author: Anthony Lopez*

Every statistical method was applied to each station for each time interval. Due to the volume of results, a subset of stations was also selected for further examination. The subset was chosen based on previous climate analysis performed by the National Climatic Data Center (NCDC). NCDC selected 9 (figure 7) climatically consistent
regions that are useful for putting current climate anomalies into a historic perspective (Karl and Koss, 1984).

Figure 7 Subset Stations

Then nine ground stations were selected that represent each climatic region. The subset station names can be seen in table 1.

Table 1 Subset Stations

<table>
<thead>
<tr>
<th>Station</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEY WEST INTL ARPT</td>
<td>FL</td>
</tr>
<tr>
<td>BROWNSVILLE S PADRE ISL INTL</td>
<td>TX</td>
</tr>
<tr>
<td>PHOENIX SKY HARBOR INTL AP</td>
<td>AZ</td>
</tr>
<tr>
<td>LOS ANGELES INTL ARPT</td>
<td>CA</td>
</tr>
<tr>
<td>WILKES-BARRE SCRANTON INTL AP</td>
<td>PA</td>
</tr>
<tr>
<td>CLEVELAND HOPKINS INTL AP</td>
<td>OH</td>
</tr>
<tr>
<td>ROCK SPRINGS ARPT [GREEN RIVER - UO]</td>
<td>WY</td>
</tr>
<tr>
<td>MINNEAPOLIS-ST PAUL INT’L ARP</td>
<td>MN</td>
</tr>
<tr>
<td>EUGENE MAHLON SWEET ARPT [UO]</td>
<td>OR</td>
</tr>
</tbody>
</table>
Ground Measurement Data

The ground measurement data came from the Integrated Surface Database (ISD), a product distributed by the National Oceanic and Atmospheric Administration’s (NOAA) National Centers for Environmental Information. ISD is a global database with over 35,000 stations of hourly observations from various sources with strict quality control measures. The ISD data used in this analysis were derived from the NSRDB Meteorological Statistical Model 3 (MTS3) as this provided a convenient and packaged source of the data. This resulted in the use of 216 ISD ground stations for statistical comparison.

Atmospheric Reanalysis Products

The three atmospheric reanalysis products evaluated were: Modern-Era Retrospective Analysis for Research and Applications (MERRA), North American Regional Reanalysis (NARR), and Climate Forecast System Reanalysis (CFSR). Each product was pre-determined to have adequate spatial and temporal resolution. It should also be noted that all products were readily available from previous efforts at NREL. They are described in detail below.

Modern-Era Retrospective Analysis for Research and Applications (MERRA)

MERRA is a National Aeronautics and Space Administration (NASA) reanalysis using the Goddard Earth Observing System Data Assimilation System Version 5. MERRA’s focus is on the historic
hydrological cycle for application in a climate context (National Aeronautics and Space Administration). MERRA has varying temporal and spatial resolution depending on the variable; however, all variables used in this analysis are hourly temporal and ~40 km$^2$ spatial resolution.

American Regional Reanalysis (NARR)
NARR is a product of the National Centers for Environmental Prediction (NCEP). NARR uses the NCEP Eta Model with the Regional Data Assimilation System (RDAS) and boasts substantial improvements in the accuracy of temperature, winds, and precipitation over NCEP-DOE Global Reanalysis 2 (Physical Sciences Division) - another reanalysis product not evaluated in this analysis. The variables in NARR have a temporal resolution of 3 hours and a spatial resolution of ~32 km$^2$.

Climate Forecast System Reanalysis (CFSR)
CFSR is a product developed by NOAA/NCEP. It was designed as a global coupled atmosphere-ocean-land surface-sea ice system with emphasis in understanding these coupled domains (Saha, 2010). The CFSR has a spatial resolution of ~38 km$^2$ and an hourly temporal resolution.

Included Variables
Each of the datasets selected for the analysis have impact on radiative transfer modeling and/or are important variables for solar energy modeling. Each variable is described below.
**Total Precipitable Water**

Total precipitable water is defined by the American Meteorological Society (Precipitable Water – AMS Glossary) as the water that is contained in a column of unit cross section extending from the earth’s surface to the top of the earth’s atmosphere. Total precipitable water is used in radiative transfer models.

**Wind Speed**

Wind speed, defined by the American Meteorological Society (Wind Speed – AMS Glossary), is the ratio of the distance covered by the air to time taken to cover it. Wind speed is modeled at varying heights; for solar resource application, the surface wind speed is used. Wind speed is an important input into solar energy modeling.

**Dew Point**

The University of Illinois defines dew point as the air temperature to which air would have to cool in order to reach saturation (Observed Dew Point Temperature). This is at constant pressure and constant water vapor content. Dew point is an input in solar energy models.

**Surface Air Temperature**

Surface air temperature is the temperature indicated by a thermometer when exposed to the air not in direct solar radiation (Air Temperature – AMS Glossary). Air temperature is used in the calculation of the solar zenith angle, an integral input into radiative
transfer models. It is also important in solar energy modeling as it impacts the efficiency of photovoltaic cells.

**Atmospheric Pressure**
Atmospheric pressure is the pressure exerted by the atmosphere resulting from gravitational attraction put upon the column of air directly above (Atmospheric Pressure – AMS Glossary). Atmospheric pressure is an input to solar zenith angle calculations, radiative transfer models, and solar energy modeling.

**Spatial Lookup**
Each of the 216 ground stations was assigned to a single grid-cell from each of the atmospheric reanalysis products. This was achieved by performing a spatial nearest neighbor lookup between the station latitude and longitude and each grid-cell centroid latitude and longitude. The cKDTree method in the SciPy Spatial Python package was used to perform the nearest neighbor lookup. The cKDTree algorithm achieves rapid performance by creating a binary tree with nodes representing an axis-aligned hyperrectangle. Each node splits the set of points into greater than or less than a particular value. This tree can then be queried to return k nearest neighbors (SciPy).

**Temporal Decimation and Averaging**
Each of the reanalysis products was first matched to the ISD ground station temporal resolution of 1 hour. MERRA and CFSR were stored with an additional ½-hour time-step developed from previous efforts.
This additional time-step was removed via decimation to produce hourly resolution. NARR was stored in its original 3-hour temporal resolution. To match the resolution of the ISD data, NARR variables were extrapolated to hourly resolution. This method of filling would follow what would have to be done in the NSRDB to achieve proper time resolution. In all datasets, leap day was removed. Each variable from both ISD ground measurement and atmospheric reanalysis products were stored in hourly resolution and also averaged by year into daily means and monthly means. This resulted in 12 values per year for monthly mean and 365 values per year for daily mean per variable. The statistical methods were then applied to each of these means.

Statistical Methods
There were five statistics methods applied to the data. RMSE, CVRMSE, MAE, and Percent Bias (all described in detail in the literature review) were applied at all 216 ISD stations for all years for both monthly and daily means. A Person product-moment correlation coefficient was applied to non-averaged data for select years on the subset of stations.
Results Storage
All data and results were stored in a single HDF. This facilitated the requirement of graphing and mapping the statistical results. The structure of the HDF can be seen in figure 8.

Challenges
There were several challenges encountered during the analysis. First was the handling of the lower resolution of the NARR data. Applying more robust methods of temporal interpolation might have resulted in better agreement with the ground measurements. There are also some questions left concerning data quality. There were several data outliers. Were these outliers a part of the original NARR dataset? Or were they introduced in NREL’s pre-processing? The same goes for some values in CFSR. Were the outliers present in the original CFSR or
were they introduced by NREL in its pre-processing? Finally, NARR, in its NREL pre-processed form, did not contain the total precipitable water variable. This was unfortunate as precipitable water is one of the driving meteorological forces in radiative transfer models. However, as will be shown below, NARR contained significant errors and thus was easily ruled out of contention for use in the NSRDB.

Results
Each variable and each statistical comparison are presented below.

Total Precipitable Water
Only the MERRA and CFSR reanalysis products contained the total precipitable water variable. CFSR had a fair agreement with ground-measured data in the Mountain West, but did poorly compared to MERRA overall (figure 9).
The Pearson’s $r$ and scatter plot (figure 10) give a false representation of the accuracy of CFSR, having an $r = 0.86$ and a scatter following the diagonal quite well while MERRA had an $r = 0.93$.

Both MERRA and CFSR show consistent RMSE accuracy across all years for the subset stations seen in figure 11.
Among all the plots, the superior accuracy of MERRA is most apparent in Figure 12, which shows the monthly percent bias across all the years. CFSR had a monthly percent bias around 90 for all months while MERRA had a monthly percent bias around 0 for all months and years.
**Wind Speed**

All three atmospheric reanalysis datasets performed similarly both in space and time for wind speed. There seemed to be a spatial pattern of higher error where there are typically higher wind speeds. This can be seen in figure 13 with higher errors in the Midwest.
CFSR and MERRA had similar Pearson’s \( r = 0.5 \) and \( r = 0.56 \) respectively, while NARR displayed an \( r = 0.035 \). Among all the variables, wind speed seemed to be the most uncertain variable among the atmospheric reanalysis products considered.
Figure 14 Wind Speed - Pearson’s R - 2009 - MINNEAPOLIS-ST PAUL INT’L ARP

All three datasets display similar spreads of RMSE across all years for the subset of stations seen in figure 15.
NARR had the smallest percent error distribution on a monthly basis across the years. Also noticeable are the higher errors in the winter months compared to the summer months for all atmospheric reanalysis products.
Dew Point
The CFSR performed the best with regards to dew point. There were a couple of outlier stations that brought its overall accuracy down, but for the most part; it matched the ground measurement data very well. The outlier stations are present in all models, which might suggest erroneous data in the ground measurements (figure 17).
MERRA performed well in the Eastern and Midwestern regions of the U.S.; however, its accuracy dropped noticeably in the West. NARR’s accuracy seemed spatially sporadic. The Pearson’s $r (r = 0.95)$ and scatter plot (figure 18) further display the agreement between CSFR and the ground measurements. MERRA displays a similar agreement with an $r = 0.9$. NARR is noticeably more spread on the scatter plot and had only an $r = 0.77$. 
All models behaved similarly over the years, maintaining roughly the same RMSE for each station (figure 19). CFSR maintained a RMSE between 1 and 5 for all subset stations across all years, while MERRA had a max RMSE of 14 and NARR a max of 26.
All models performed better in the summer months than in the winter months as seen in boxplots in figure 20. The distribution of monthly percent bias errors is largest in NARR followed by MERRA then CFSR. However, there are large outliers in all, which is a result of the outlier stations.
Surface Temperature
Not surprisingly the surface temperature variable displays the same
spatial and temporal errors as dew point temperature for all
atmospheric reanalysis products. This is a result of dew point
temperatures intrinsic link with air temperature and water vapor.
There remain the outlier stations, again, most likely a result of errors
in the ground measurement data (figure 21). The CFSR agreed with
the ground measurements better than MERRA and NARR across the
United States.
Both CFSR and MERRA displayed accurate Pearson’s R ($r = 0.94$) and agreement in the scatter plots. NARR fared far worse with an $r = 0.76$. Interestingly, as seen in the histograms in figure 22, CFSR and ISD display bimodal distributions while MERRA and NARR are closer to normal distributions for the particular station graphed.
All atmospheric reanalysis products displayed similar RMSE across the years for the subset stations. NARR had the largest range of RMSE (3 to 21) while CFSR and MERRA had a smaller range of RMSE between 1 and 4.
The surface temperature percent bias shows a similar monthly pattern as dew point with higher errors in the winter months and lower in the summer months (figure 24). NARR has the largest range of errors but all exhibit substation outliers, a result of the outlier stations.
Atmospheric Pressure
CFSR and MERRA display a similar spatial pattern in atmospheric pressure MAE errors - performing better in the Eastern and Midwestern United States with larger errors in the Western United States (figure 25). NARR had significant MAE in all regions of the United States.
Both CFSR and NARR have peculiar errors in atmospheric pressure as seen in figure 26. There is some question as to whether these are true errors or a result of NREL pre-processing. MERRA shows great agreement with the ground measurements with an $r = 0.94$. 
MERRA has similar RMSE across all years while NARR and CFSR exhibit larger RMSE in 2008 and 2009 respectively when compared to all other years RMSE (figure 27). NARR seemed to have particular trouble with its accuracy for one of the subset stations while CFSR and MERRA performed similarly for all subset stations.
Figure 27 Atmospheric Pressure - RMSE - All Years - Station Subset

MERRA and NARR had consistent percent bias error across all months for all years for the subset stations while CFSR had higher errors in March, April, and September (figure 28). NARR exhibited large outliers in all months, most likely associated with a single station in the subset.
Discussion and Summary

The relative model accuracy of the select variables from three atmospheric reanalysis products was evaluated against measured data. The evaluation process included five different statistical measures: root mean square error, coefficient of variation of the root mean square error, Pearson’s r, percent bias, and mean absolute error. Each statistical method was applied at each ISD station location and its associated atmospheric reanalysis grid-cell. The associated grid-cells were assigned using a spatial nearest neighbor lookup. Each statistical method was performed for years 2005-2009 to daily and
monthly means revealing seasonal errors. Maps were produced for the MAE statistic that revealed spatial trends in errors. Each atmospheric reanalysis variable bias was then described qualitatively. It was found that NARR consistently had the most errors for all variables in both space and time. CFSR outperformed MERRA in dew point; however, MERRA outperformed in total precipitable water and atmospheric pressure. CFSR and MERRA performed similarly in surface temperature. Given the greater importance and influence of atmospheric pressure and more so, total precipitable water on radiative transfer models, MERRA seems a likely choice for inclusion in the NSRDB. It must be noted, however, there were some unexplained outliers in the ISD ground measurements that could have been a result of NREL pre-processing and thus would have influence on the statistical measures. It should also be noted that there were also unexplained outliers in both CFSR and NARR that also could have been a result of NREL pre-processing. These should be investigated further to avoid any mischaracterization. Still, the overall accuracy of MERRA over CFSR and NARR in the select variables is large enough to recommend its use in the NSRDB. Further research is also suggested to quantify the impacts of using different atmospheric reanalysis products in solar resource assessment to better understand the importance of using accurate meteorological variables.
Appendix

Analysis Python Code

```python
% matplotlib inline
import h5py
import numpy as np
import pandas as pd
import time
from scipy.spatial import cKDTree
from matplotlib import pyplot as plt
plt.style.use('ggplot')

# These have all years of data in one file
isd = h5py.File('/scratch/alopez/metcompare/nsrdb_9110.h5', 'r')
narr = h5py.File('/scratch/alopez/metcompare/narr_timeseries.h5', 'r')

# These have their years of data separated
merra_path = '/scratch/alopez/blend/v2.0.1/'
cfsr_path = '/projects/PXS/nsrdb/v1.0.2/'

# ISD Indices to Analyze
isd_csv = pd.read_csv('/scratch/alopez/metcompare/isd_stations.csv')
isd_meta = pd.DataFrame(isd['meta']['usaf'])
isd_meta['h_index'] = isd_meta.index
isd_meta.columns = ['usaf', 'h_index']

# these are the proper indices to analyze, only includes true ISD locations
isd_index = pd.merge(isd_meta, isd_csv, how='inner')['h_index']

# dictionary for easy name lookup
vars = {'precip': {'isd': 'prec_wat', 'merra': 'total_precipitable_water', 'narr': 'na', 'cfsr': 'total_precipitable_water_nwp'},
        'temp': {'isd': 'temp_dryb', 'merra': 'surface_temperature', 'narr': 'temperature', 'cfsr': 'surface_air_temperature_nwp'},
        'dew_point': {'isd': 'dew_point', 'merra': 'dew_point', 'narr': 'dew_point', 'cfsr': 'dew_point'},
        'wspd': {'isd': 'wind_spd', 'merra': 'wspd', 'narr': 'wind_speed', 'cfsr': 'wind_speed_10m_nwp'},
        'pres': {'isd': 'atm_pres', 'merra': 'surface_pressure', 'narr': 'pressure', 'cfsr': 'surface_pressure_background'}}

# Take a look at the units
print('ISD Units')
print(isd['prec_wat'].attrs['units'])
print(isd['temp_dryb'].attrs['units'])
print(isd['dew_point'].attrs['units'])
print(isd['wind_spd'].attrs['units'])
print(isd['atm_pres'].attrs['units'])
print('')
print('MERRA Units')
with h5py.File(merra_path + 'nsrdb_2005.h5', 'r') as merra:
    print(merra['total_precipitable_water'].attrs['units'])
    print(merra['surface_temperature'].attrs['units'])
    print(merra['dew_point'].attrs['units'])
    print(merra['wspd'].attrs['units'])
    print(merra['surface_pressure'].attrs['units'])
print('')
print('CFSR Units')
with h5py.File(cfsr_path + 'nsrdb_2005.h5', 'r') as cfsr:
    print(cfsr['total_precipitable_water_nwp'].attrs['units'])
    print(cfsr['surface_air_temperature_nwp'].attrs['units'])
    print(cfsr['dew_point'].attrs['units'])
    print(cfsr['wind_speed_10m_nwp'].attrs['units'])
    print(cfsr['surface_pressure_background'].attrs['units'])
print('')
print('NARR Units')
print(narr['temperature'].attrs['unit'])
```

In [1]:

In [2]:

In [3]:

In [4]:

print(narr['dew_point'].attrs['unit'])
print(narr['wind_speed'].attrs['unit'])
print(narr['pressure'].attrs['unit'])

ISD Units
cm
°C
m/s
mbar

MERRA Units
cm
Kelvin
Kelvin
m/s
mbar

CF SR Units
cm
Kelvin
Kelvin
m/s
mbar

NARR Units
degrees C
degrees C
meters/second
millibars

# Perform the spatial lookup for each station and reanalysis dataset
isd_coords = np.dstack((isd['meta']['nsrdb_lon'][isd_index].ravel(),
isd['meta']['nsrdb_lat'][isd_index].ravel()))[0]
narr_coords = np.dstack((narr['meta']['lng'][...].ravel(),
narr['meta']['lat'][...].ravel()))[0]
with h5py.File(merra_path + 'nsrdb_2005.h5', 'r') as merra:
    merra_coords = np.dstack((merra['meta']['longitude'][...
    merra['meta']['latitude'][...].ravel()))[0]
with h5py.File(cfsr_path + 'nsrdb_2005.h5', 'r') as cfsr:
    cfsr_coords = np.dstack((cfsr['meta']['longitude'][...

# Create the output HDF
hfile = h5py.File('/scratch/alopez/metcompare/results.h5', 'w')
meta = np.dtype([('latitude', np.float), ('longitude', np.float),
('station', 'S30'), ('state', 'S10'), ('usaf', 'S30'),
('merra_latitude', np.float), ('merra_longitude', np.float),
('narr_latitude', np.float), ('narr_longitude', np.float),
('cfsr_latitude', np.float), ('cfsr_longitude', np.float),
('merra_index', int), ('narr_index', int), ('cfsr_index', int)])
dset = hfile.create_dataset('meta', shape=(isd_index.shape[0],),
dtype=meta)
dset['longitude'] = isd['meta']['nsrdb_lon'][isd_index].ravel()
dset['latitude'] = isd['meta']['nsrdb_lat'][isd_index].ravel()
dset['station'] = isd['meta']['station'][isd_index].ravel()
dset['state'] = isd['meta']['state'][isd_index].ravel()
dset['usaf'] = isd['meta']['usaf'][isd_index].ravel()

dset[narr['narr_longitude']] = narr_coords[narr_index][:, 0]
for var in vars.iterkeys():
    hfile.create_group(var)
for year in range(2005, 2010):
    hfile.create_group(f'{v}/{y}').format(v=var, y=year)
    hfile.create_group(f'{v}/{y}/data').format(v=var, y=year)
    hfile.create_group(f'{v}/{y}/results').format(v=var, y=year)

# iterating each ISD station - vectorized operations would make this analysis trivial, however,
# there are several duplicated target indices (resulting from ISD stations being spatially close)
# which make extraction from HDF a pain, even with fancy indexing...
isd_ti = pd.to_datetime(isd['time_index'][:, :].ravel())
for i, isd_i in enumerate(isd_index):
    for var in vars.iterkeys():
        for year in range(2005, 2010):
            tic = time.time()
            # get ISD data first
            isd_data = pd.DataFrame(isd[vars[var][isd_i][isd_i.year == year, isd_i]]
            index = isd Ti[isd_i.year == year]
            # excluding leap day from analysis... might not actually be needed
            isd_data = isd_data[~((isd_data.index.month == 2) & (isd_data.index.day == 29))]
            isd_month = isd_data.groupby([(isd_data.index.month), .mean().values.ravel()]
            isd_day = isd_data.groupby([(isd_data.index.month, isd_data.index.day)]).mean().values.ravel()

            # get narr data next (does not have precip)
            if not var == 'precip':
                narr_data = pd.DataFrame(narr[vars[var][narr_i][narr_i.year == year, narr_i]]
                index = narr Ti[narr_i.year == year] * 0.001
                narr_data = narr_data[~((narr_data.index.month == 2) & (narr_data.index.day == 29))]
                narr_month = narr_data.groupby([(narr_data.index.month), .mean().values.ravel()]
                narr_day = narr_data.groupby([(narr_data.index.month, narr_data.index.day)]).mean().values.ravel()

            with h5py.File(merra_path + f'nsrdb_{y}.h5').format(y=year, 'r') as merra:
                # get MERRA. Merra is originally hourly, however data is stored in 1/2-hourly, so removing dups
                merra_ti = pd.to_datetime(merra['time_index'][:, :].ravel())
                merra_data = pd.DataFrame(merra[vars[var][merra_i][merra_i.year == year, merra_i]]
                index = merra Ti[merra_i.year == year] * 0.001
                merra_data = merra_data[~((merra_data.index.month == 2) & (merra_data.index.day == 29))]
                merra_month = merra_data.groupby([(merra_data.index.month), .mean().values.ravel()]
                merra_day = merra_data.groupby([(merra_data.index.month, merra_data.index.day)]).mean().values.ravel()

            with h5py.File(cfsr_path + f'nsrdb_{y}.h5').format(y=year, 'r') as cfsr:
                # get CFSR. CFSR is originally hourly, however data is stored in 1/2-hourly, so removing dups
                cfsr_ti = pd.to_datetime(cfsr['time_index'][:, :].ravel())
                cfsr_data = pd.DataFrame(cfsr[vars[var][cfsr_i][cfsr_i.year == year, cfsr_i]]
                index = cfsr Ti[cfsr_i.year == year] * 0.001
                cfsr_data = cfsr_data[~((cfsr_data.index.month == 2) & (cfsr_data.index.day == 29))]
                cfsr_month = cfsr_data.groupby([(cfsr_data.index.month), .mean().values.ravel()]
                cfsr_day = cfsr_data.groupby([(cfsr_data.index.month, cfsr_data.index.day)]).mean().values.ravel()
cfsr_month = cfsr_month - 273.15

cfsr_day = cfsr_day - 273.15
merra_month = merra_month - 273.15
merra_day = merra_day - 273.15

# perform statistics
# RMSE ---------------------
if not var = 'precip':
    narr_rmse_day = np.sqrt(((narr_day - isd_day) ** 2).mean())
    narr_rmse_month = np.sqrt(((narr_month - isd_month) ** 2).mean())

merra_rmse_day = np.sqrt(((merra_day - isd_day) ** 2).mean())
merra_rmse_month = np.sqrt(((merra_month - isd_month) ** 2).mean())
cfsr_rmse_day = np.sqrt(((cfsr_day - isd_day) ** 2).mean())
cfsr_rmse_month = np.sqrt(((cfsr_month - isd_month) ** 2).mean())

# CVRMSE ---------------------
if not var = 'precip':
    narr_cvrms_day = narr_rmse_day / isd_day * 100.
    narr_cvrms_month = narr_rmse_month / isd_month * 100.

merra_cvrms_day = merra_rmse_day / isd_day * 100.
merra_cvrms_month = merra_rmse_month / isd_month * 100.
cfsr_cvrms_day = cfsr_rmse_day / isd_day * 100.
cfsr_cvrms_month = cfsr_rmse_month / isd_month * 100.

# Percent Bias
if not var = 'precip':
    narr_pb_day = (isd_day - narr_day) / isd_day * 100.
    narr_pb_month = (isd_month - narr_month) / isd_month * 100.

merra_pb_day = (isd_day - merra_day) / isd_day * 100.
merra_pb_month = (isd_month - merra_month) / isd_month * 100.
cfsr_pb_day = (isd_day - cfsr_day) / isd_day * 100.
cfsr_pb_month = (isd_month - cfsr_month) / isd_month * 100.

# MAE
if not var = 'precip':
    narr_mae_day = np.sum(np.abs(isd_day - narr_day)) / isd_day.shape[0]
    narr_mae_month = np.sum(np.abs(isd_month - narr_month)) / isd_month.shape[0]

merra_mae_day = np.sum(np.abs(isd_day - merra_day)) / isd_day.shape[0]
merra_mae_month = np.sum(np.abs(isd_month - merra_month)) / isd_month.shape[0]
cfsr_mae_day = np.sum(np.abs(isd_day - cfsr_day)) / isd_day.shape[0]
cfsr_mae_month = np.sum(np.abs(isd_month - cfsr_month)) / isd_month.shape[0]

# create datasets if first index
if i = 0:
    hfile.create_dataset('[v]/[y]/data/isd_data'.format(v=var, y=year),
        shape=(isd_data.shape[0], isd_index.shape[0]), dtype=pa.float)

if not var = 'precip':
    hfile.create_dataset('[v]/[y]/data/narr_data'.format(v=var, y=year),
        shape=(narr_data.shape[0], isd_index.shape[0]), dtype=pa.float)
    hfile.create_dataset('[v]/[y]/data/merra_data'.format(v=var, y=year),
        shape=(merra_data.shape[0], isd_index.shape[0]), dtype=pa.float)
    hfile.create_dataset('[v]/[y]/data/cfsr_data'.format(v=var, y=year),
        shape=(cfsr_data.shape[0], isd_index.shape[0]), dtype=pa.float)

if not var = 'precip':
    hfile.create_dataset('[v]/[y]/results/narr_rmse_day'.format(v=var, y=year), shape=(1,
        isd_index.shape[0]), dtype=pa.float)
    hfile.create_dataset('[v]/[y]/results/narr_rmse_month'.format(v=var, y=year), shape=(1,
# store the data
hfile['{v}/{y}/data/isd_data'.format(v=var, y=year)][i] = isd_data.values.ravel()
if not var == 'precip':
    hfile['{v}/{y}/data/narr_data'.format(v=var, y=year)][i] = narr_data.values.ravel()
    hfile['{v}/{y}/data/merra_data'.format(v=var, y=year)][i] = merra_data.values.ravel()
    hfile['{v}/{y}/data/cfsr_data'.format(v=var, y=year)][i] = cfsr_data.values.ravel()

# store the results
if not var == 'precip':
    hfile['{v}/{y}/results/narr_rmse_day'.format(v=var, y=year)][i] = narr_rmse_day
    hfile['{v}/{y}/results/narr_rmse_month'.format(v=var, y=year)][i] = narr_rmse_month
    hfile['{v}/{y}/results/narr_crms_day'.format(v=var, y=year)][i] = narr_crms_day
    hfile['{v}/{y}/results/narr_crms_month'.format(v=var, y=year)][i] = narr_crms_month
    hfile['{v}/{y}/results/narr_pb_day'.format(v=var, y=year)][i] = narr_pb_day
    hfile['{v}/{y}/results/narr_pb_month'.format(v=var, y=year)][i] = narr_pb_month
    hfile['{v}/{y}/results/narr_mae_day'.format(v=var, y=year)][i] = narr_mae_day
    hfile['{v}/{y}/results/narr_mae_month'.format(v=var, y=year)][i] = narr_mae_month

hfile.close()
hfile['{v}/{y}/results/merra_rmse_day'.format(v=var, y=year)][0, i] = merra_rmse_day
hfile['{v}/{y}/results/merra_rmse_month'.format(v=var, y=year)][0, i] = merra_rmse_month
hfile['{v}/{y}/results/merra_cvrmse_day'.format(v=var, y=year)][0, i] = merra_cvrmse_day
hfile['{v}/{y}/results/merra_cvrmse_month'.format(v=var, y=year)][0, i] = merra_cvrmse_month
hfile['{v}/{y}/results/merra_pb_day'.format(v=var, y=year)][0, i] = merra_pb_day
hfile['{v}/{y}/results/merra_pb_month'.format(v=var, y=year)][0, i] = merra_pb_month

hfile['{v}/{y}/results/cfsr_rmse_day'.format(v=var, y=year)][0, i] = cfsr_rmse_day
hfile['{v}/{y}/results/cfsr_rmse_month'.format(v=var, y=year)][0, i] = cfsr_rmse_month
hfile['{v}/{y}/results/cfsr_cvrmse_day'.format(v=var, y=year)][0, i] = cfsr_cvrmse_day
hfile['{v}/{y}/results/cfsr_cvrmse_month'.format(v=var, y=year)][0, i] = cfsr_cvrmse_month
hfile['{v}/{y}/results/cfsr_pb_day'.format(v=var, y=year)][0, i] = cfsr_pb_day
hfile['{v}/{y}/results/cfsr_pb_month'.format(v=var, y=year)][0, i] = cfsr_pb_month
hfile['{v}/{y}/results/cfsr_mae_day'.format(v=var, y=year)][0, i] = cfsr_mae_day
hfile['{v}/{y}/results/cfsr_mae_month'.format(v=var, y=year)][0, i] = cfsr_mae_month

hfile.close()
References


Chai, T., and R. R. Draxler. "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? - Arguments Against Avoiding RMSE in the Literature". Geoscientific Model Development: 1247-1250.


